ENSEMBLE MACHINE LEARNING MODEL GENERALIZABILITY AND ITS APPLICATION TO INDIRECT TOOL CONDITION MONITORING

A Thesis
Presented to
The Academic Faculty

By

Alexandra G. Schueller

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in the
George W. Woodruff School of Mechanical Engineering

Georgia Institute of Technology

December 2021

ENSEMBLE MACHINE LEARNING GENERALIZABILITY AND ITS APPLICATION TO INDIRECT TOOL CONDITION MONITORING

Approved by:

Dr. Christopher Saldana, Advisor
George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Dr. Katherine Fu, Co-Advisor
George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Dr. Thomas Kurfess, Co-Advisor
George W. Woodruff School of Mechanical Engineering
Georgia Institute of Technology

Date Approved: August 16th, 2021

ACKNOWLEDGEMENTS

I would like to thank my advisor here at Georgia Tech, Dr. Christopher Saldana, for his support, insight, and guidance throughout this project. You have helped me grow as a researcher and develop new ways of thinking about problems and opportunities over the last two years. I would also like to thank my committee members and co-advisors, Drs. Katherine Fu and Thomas Kurfess, for their constant enthusiasm about our team's projects, their invaluable insights, and their willingness to help with anything they can. You have both served as great mentors and sources of positivity throughout this process.

I would also like to thank my EPICS teammates for their encouragement, humor, friendships, and help with any questions I've had.

Thank you to my family for always supporting me, believing in me, and helping me navigate my career thus far. Finally, I would like to thank my fiancé, MJ, for his love, support, and patience throughout graduate school and everything that has come with it.

This work was supported in part by the Department of Energy DE-EE0008303.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

IIoT	Industrial Internet of Things
TCM	Tool condition monitoring
ML	Machine learning
DT	Decision tree
SVM	Support vector machine
kNN	K-nearest neighbors
ANN	Artificial neural network
RF	Random forest
ET	Extremely randomized trees (“extra trees”)
AWGN	Additive white Gaussian noise
CV	Cross validation
LOGO-CV	Leave-one-group-out cross validation
CNC	Computer numerically controlled
VB	Flank wear
VB_{max}	Maximum flank wear
AE	Acoustic emission
DFT	Discrete Fourier Transform
TD	Time domain
FD	Frequency domain
EHV	Ensemble Hard Voting
ESV	Ensemble Soft Voting
VB	Flank wear
RFE	Recursive Feature Elimination
RFECV	RFE using cross validation
SVM-RFECV	RFE using CV and an SVM model
RF-RFECV	RFE using CV and a RF model
MI	Mutual Information
CFS	Correlation-based feature selection
T	Spindle rotation period
SP	Spindle power
XL	X-axis load
YL	Y-axis load
ZL	Z-axis load
t	Machining time
MR	Material removed
N	Spindle speed
FR	Feed rate
RMS	Root-mean-square
FS	Feature selection
SD	Standard deviation

SUMMARY

A practical, accurate, robust, and generalizable system for monitoring tool condition during a machining process would enable advancements in manufacturing process automation, cost reduction, and efficiency improvement. Previously proposed systems using various individual machine learning (ML) models and other analysis techniques have struggled with low generalizability to new machining and environmental conditions, as well as a common reliance on expensive or intrusive sensory equipment which hinders their industry adoption. While ensemble ML techniques offer significant advantages over individual models in terms of performance, overfitting reduction, and generalizability improvement, they have only begun to see limited applications within the field of tool condition monitoring (TCM).

To address the research gaps which currently surround TCM system generalizability and optimal ensemble model configuration for this application, nine ML model types, including five heterogeneous and homogeneous ensemble models, are employed for tool wear classification. Sound, spindle power, and axial load signals are utilized through the sensor fusion of practical external and internal machine sensors. This original experimental process data is collected through tool wear experiments using a variety of machining conditions. Four feature selection methods and multiple tool wear classification resolution values are compared for this application, and the performance of the ML models is compared across metrics including k-fold cross validation and leave-one-group-out cross validation. The generalizability of the models to data from unseen experiments and machining conditions is evaluated, and a method of improving the generalizability levels using noisy training data is examined. T-tests are used to measure the significance of model performance differences. The extra-trees ensemble ML method, which had never before been applied to signal-based TCM, shows the best performance of the nine models.

CHAPTER 1: INTRODUCTION

Amid constant industry demand for manufacturers to increase their production and efficiency, many businesses are turning to automation, cyber-physical systems, and cloud computing methods to remain competitive. This trend, known as Industry 4.0 or the fourth industrial revolution, requires the exchange of increasing amounts of data in order for advantageous process decisions to be made automatically. Industrial Internet of Things (IIoT), the technology which allows production data from a network of connected sensors and machines to be monitored continuously, can be used to gain valuable insight into manufacturing process conditions, issues, and efficiency improvement opportunities.

1.1 Motivation

Tool condition monitoring (TCM) has been an area of interest for decades due to demands like these [1-8]. For a variety of metal manufacturing processes, worn tools are known to result in parts with poor surface finish, dimensional inaccuracy, decreased fatigue strength, and altered material properties [1, 2, 9-13]. The continued use of worn tools also increases power consumption, process temperatures, and machine vibration [12]. Finally, tools at the end of their lives can cause expensive damage to machines and workpieces, unexpected downtime for repairs, and danger to machine operators, when they eventually fail catastrophically [3, 6, 8, 12, 14-16]. Tool failure alone has been reported to cause 7 – 20% [8, 17] of machine downtime.

Currently, businesses attempt to avoid these negative effects by changing their tools out after a very conservative number of parts set by the machine operator [7, 9, 18-20]. However, this practice also has negative consequences: it decreases productivity and significantly increases the

overall tooling cost [2]. For the milling operation, for example, tools and tool changes have been found to account for 3-12% of the total processing cost [21]. In addition, individual tools' lifetimes can vary greatly due the complexity of machining processes and a variety of relatively random factors [20]. As a result, some tools may still fail early even with conservative practices employed [19]. Some experienced operators are able to get an idea of when a tool is nearing the end of its life by identifying changes in its cutting sound [22-24]. However, this technique is highly subjective and is a roadblock for increased process automation. If instead, a tool's wear level could be reliably and automatically monitored during the manufacturing process, tools could be used for longer periods while avoiding detrimental wear levels, fewer tools would need to be purchased, manufacturing productivity could be increased, downtime could be decreased, and machining costs could be reduced by as much as 10-40% [3, 9, 11, 25, 26].

1.2 Gaps in Research

While several research groups have studied this issue and employed various methods to address turning, milling, drilling, and grinding processes, a practical and reliable TCM system for industry machining has yet to be offered [5, 7, 19, 27, 28]. This is due to several factors, including the high levels of noise and vibration on shop floors; the wide range of cutting conditions used for industrial production; machining processes' time-variant nature; the nonlinear relationships between measured features and tool wear; and laboratory sensory equipment that is too expensive or intrusive to be implemented commercially [19, 22, 23, 28-31]. While it is recognized that previously developed systems have generally performed poorly in production environments like these, the generalizability of TCM systems is not well understood [19, 32]. Specifically, the generalizability of TCM classification models to data from unseen experiments with specific

machining condition changes has not been evaluated or compared. In addition, the "noisy training" technique, which has been shown to improve neural network generalizability to new conditions in other fields [33-35], has never been applied to TCM applications.

The methods applied to TCM have varied greatly, ranging from the use of tool life equations [36-39], remaining useful life predictions [26, 40, 41], and process signal frequency spectrum analysis [13, 14, 22, 29, 42-44], to the more recent application of machine learning (ML) techniques [2, 4, 5, 12, 18, 22, 29, 45-47]. While different machine learning model types, such as decision trees, support vector machines, k-nearest neighbors, and artificial neural networks have presented various advantages for TCM, it remains difficult for any one model to achieve high classification accuracies while also avoiding over-fitting and keeping generalization ability high [28-48]. Ensemble machine learning, a technique which combines multiple base models to create an improved final model, has recently been shown in limited cases to improve tool wear classification performance through the analysis of surface texture images [49], spindle motor current and power [14], vibration [50], acoustic emission [51], and cutting force [48, 52]. However, despite several researchers' findings showing significant feasibility and performance advantages of sound signals over other process data types for machine learning TCM [2, 3, 22, 23, 29, 30, 43, 53], sound's effectiveness when combined with ensemble ML techniques has yet to be studied. In addition, certain ensemble ML types such as the soft ensemble voting and extremely randomized trees methods have never been studied for signal-based TCM.

Sensor fusion, a technique in which the signals and unique patterns from multiple types of sensors are utilized within one algorithm, has also been shown to improve machine learning model performance and generalization ability for TCM [2, 4, 5, 7, 28, 32, 53]. However, it has rarely been combined with the advantages of ensemble ML for tool wear classification and often relies on

expensive or intrusive sensors [6, 13, 54]. Specifically, the performance of an ensemble machine learning TCM model utilizing the fusion of sound signals with data from other sensors has never been evaluated.

1.3 Problem Statement

In order to fill the research gaps surrounding TCM model generalizability, ensemble ML techniques, and sensor fusion using practical sensors, this study explores new tool wear classification methods through the application of various ensemble machine learning techniques, cost-effective and non-intrusive sensors, and methods of model generalizability improvement. Through this, the objectives of the study are to understand how the configuration of ensemble machine learning models affect their classification performance for TCM, as well as how variability in machining and environmental conditions may impact or be used to improve the generalizability of TCM models like these. To accomplish this, end milling experiments are conducted in which sound, spindle power, 3-dimensional axial load, and tool wear images are collected and analyzed along the entire useful lives of 8 tools.

Machine learning techniques are employed to classify the tool's current wear condition between three wear levels. While nine classification machine learning models were evaluated, three overall ML techniques were studied: individual or "base" models, heterogeneous ensemble models, and homogeneous ensemble models. The four base models evaluated were the decision tree, support vector machine, k-nearest neighbors, and artificial neural network algorithms. The three heterogeneous ensemble models; hard voting, soft voting, and stacked generalization techniques; were each built upon the four base models. Finally, the two homogeneous ensemble

models, the random forest and extremely randomized trees (or "extra-trees") techniques, were both formed from collections of decision trees.

The generalizability of the models to new machining conditions and specific condition changes was evaluated using experiments of various spindle speeds and feed rates. To identify if the noisy training technique can increase model generalizability for TCM and the ML algorithms investigated, the results of models trained on various levels and combinations of additive White Gaussian noise were assessed. Several performance metrics including repeated 10-fold cross validation and Leave-One-Group-Out cross validation (LOGO-CV) were also employed to quantify the stability and reliability of the models, as well as to prevent over-fitting. Finally, statistical t-tests and 95% confidence intervals were used to assess the significance of model performance differences.

1.4 Structure

The thesis paper is organized as follows: In Chapter 2, the background research in the tool condition monitoring field will be discussed with respect to tool wear patterns, the signals used for tool wear classification, and the conventional and machine learning analysis methods which have been applied to TCM. Chapter 3 will then focus on the experimental and analysis methods used for this study. Chapter 4 will assess the initial experimental, feature selection, and machine learning results. Chapter 5 will discuss the main impacts of the research, including the comparison of model performance across individual and ensemble ML types, the effect of tool wear resolution on model performance, the study of machining condition changes' effect on model performance and generalizability, and the impact of noisy training on TCM system generalizability to data from

new experiments. Finally, in Chapter 6, the main contributions of the study will be presented, its assumptions and limitations will be discussed, and areas for future research will be identified.

CHAPTER 2: BACKGROUND

The recognized need within the manufacturing industry for a practical, reliable, and accurate tool condition monitoring system has driven efforts across a wide range of expertise areas to make this a reality. Several sensors, data streams, and analysis methods have been designed, tested, and built upon to make up the body of research currently available. This chapter reviews the data types and analysis frameworks which have contributed to the current state of TCM research.

2.1 Tool Wear Progression

While several types of wear can occur during machining processes, including flank wear, face wear, crater wear, chipping, and cracking [55], flank wear (VB) has been found to be the predominant wear mode for commonly used conditions, as well as the mode which most highly affects final product quality [19, 22, 56]. As a result, it is considered the most commonly accepted measure of tool wear [2, 22, 55]. As shown for an end mill in Figure 2.1 from tool life testing standard ISO8688-2 [55], the flank wear is measured as the distance from the tool's original cutting edge to the end of the abrasive wear region on the flank surface.

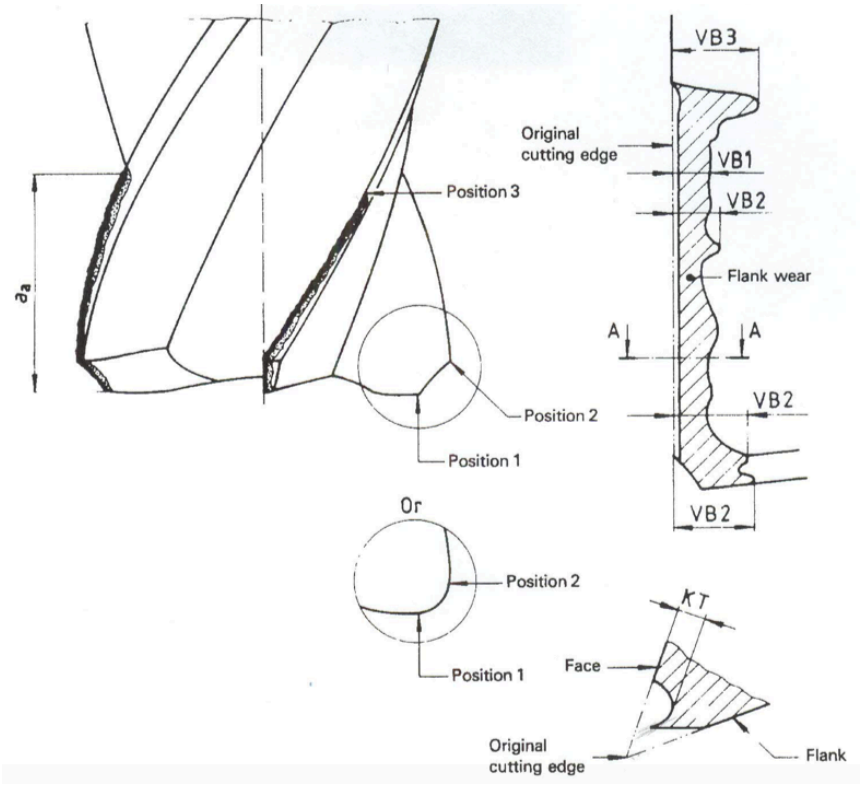


Figure 2.1: End mill flank wear [55]. Uniform flank wear (VB1), non-uniform flank wear (VB2), and localized flank wear (VB3) are shown.

This type of wear occurs gradually due to the friction between the workpiece and the flank surface of the tool, which causes small particles of the tool's material to adhere to the workpiece and then be sheared off [56].

It has been established that a machining tool's flank wear generally follows a path of three stages: the initial wear stage, the normal wear stage, and the severe wear stage [57-59]. During the relatively short initial wear stage, also known as the break-in stage, the tool's cutting edge radius changes quickly and a significant portion of any outer layer is worn away. Then, during the normal or steady state wear stage, the tool wears at a fairly constant rate for a longer period of time. Finally, a level of wear is reached which makes the process no longer sustainable and the severe wear stage is entered. During this wear stage, the tool's wear accelerates rapidly until catastrophic

failure eventually occurs. While it is the level, or measurement, of tool wear which most directly affects the production quality and is generally used to determine optimal tool change times, an understanding of the wear stages provides insight into the process mechanics [2, 19, 56].

2.2 Signals used for TCM

Many types of sensors and data have been applied to the tool condition monitoring effort over the last several decades. In TCM, these methods used to determine tool wear are generally categorized into "direct" and "indirect" measurement techniques. In order for many early TCM methods to be satisfactory, they relied on direct techniques which require interrupting the machining process to measure the wear itself. These techniques, such as visual inspection, optical sensors, workpiece size measurements, electrical resistance, and radioactive isotopes, decrease productivity significantly, often require a tool's removal from the machine, and may be negatively affected by common conditions like the presence of coolant, cutting chips, and built-up edges [7, 14, 28, 43]. This generally makes them impractical for production applications [5, 6, 12, 14, 22, 28, 43].

Indirect measurement techniques, on the other hand, can be carried out without stopping the operation, and can be lower cost and less intrusive on machining processes [5, 6, 12, 22, 50, 60]. However, as indirect techniques measure signals that are related to the tool wear instead of the wear itself, they generally require high precision and stability from the analysis algorithms [50]. Commonly studied indirect process signals include cutting force, motor current and power, vibration, acoustic emission, and sound [6, 7, 12, 14, 52].

2.2.1 Cutting Forces

Machining forces, including the main cutting force as well as feed and axial forces, have been shown to have strong correlations to tool wear due to the increase in contact surface and abrasion between the tool and the workpiece as a tool's cutting edge becomes dulled [7, 30, 61-65]. These force signals are frequently chosen for research studies due to the signals' high levels of stability and sensitivity to the wear level [40, 66]. Cutting forces are often measured during cutting processes using a sensor called a dynamometer [6, 15, 30, 40-42, 60, 67-68]. These sensors have several advantages including high measurement accuracy, robustness, and fast response time [40]. However, they are also expensive, invasive, difficult to mount, have limited frequency ranges, and cause a reduction of the local machine stiffness [19, 20, 28]. These drawbacks unfortunately make their application to production settings relatively impractical [20, 29, 31].

2.2.2 Motor Current & Power

A few alternative methods of measuring or estimating the process forces more discreetly and efficiently have gained attention for these reasons. Over the years, several research groups have effectively measured axial cutting forces indirectly through the use of feed drive motor current sensors, which are much lower-cost and easier to install than dynamometers [30, 69, 70]. These signals are directly dependent on a process' cutting forces because as a tool wears and its cutting forces increase, more power, and therefore more current, is drawn to the spindle and axial drive motors in order to continue the machining process [12]. Ghosh et al. [27] confirmed this by showing that for their face milling experiments, the measured spindle power closely follows the same pattern as the cutting force. As a result, these current sensors can serve as more cost-effective

and practical alternatives to direct force sensors with only a slight decrease in sensitivity to tool wear [7, 12, 28, 71-73].

In addition to these advantages, motor current and power sensors are increasingly being included in computer numerically controlled (CNC) machines for process feedback, quality control, and error debugging [31]. This development also allows these signals to be utilized by researchers without additional sensor installation or intrusion on the cutting process. For example, Farias et al. [16] employed these convenient machine control sensors to address TCM through the use of the communication protocol MTConnect. By monitoring the machine's spindle motor load, the x-axis motor load, the z-axis motor load, and the spindle power signals, their team was able to determine on a binary basis if flank wear levels of 0 mm or 0.3 mm were present. While their study was limited by a low sampling rate of 1 Hz and the lack of intermediate wear stages, this convenient measurement system showed promising advantages for busy and cost-conscious production environments. Through the use of the similar MQTT IoT messaging protocol, Xi et al. [20] also avoided the use of expensive equipment for their TCM study by estimating their process' cutting forces using the measured tool position values from both motor and linear encoders present within their CNC machine, as well as calculated system stiffness characteristics. Their team's work shows a strong and clear correlation between the axial load in the z-direction and the level of tool wear present.

2.2.3 Vibration

Vibration is another signal which has been commonly used for TCM applications in past research. During machining processes, cyclic movements like tool or workpiece rotations cause vibration which can decrease part quality, increase tool wear, and produce excessive sound [12].

However, these signals also show a high sensitivity to changes in tool wear, and can be used to monitor it [20, 74]. Unlike dynamometers, accelerometers are generally very cost effective and easy to install within a machine [14, 60]. The downsides to vibration-based TCM are that they are very sensitive to their placement location, as well as subject to high levels of signal noise due to toolpath changes, entering or exiting a workpiece, chip fracture, use of coolant, and general machine vibration [14, 20].

2.2.4 Acoustic Emission

Acoustic emission (AE), or the elastic stress waves created through a material during the rearrangement of its internal structure as a tool passes through it, is also a common measurement technique used for TCM [2, 7, 12, 13, 22, 51]. This signal can provide useful information about a wide variety of process conditions including tool wear, and shows a higher frequency range than those of vibration and sound [12, 23]. However, it has the disadvantages that it is more sensitive to signal noise and machining conditions than it is to the tool condition, leading some to say that it is not suitable for use on its own, as well as being fairly intrusive for the process [23].

2.2.5 Sound

Sound is also a promising signal to be used for TCM, although it has been studied to a lesser extent by the research community [23, 29, 43]. The sound signals are created by friction between the cutting tool and the workpiece during cutting, and therefore are significantly affected by the dulling of the tool's edge and the resulting increase in contact surface area when a tool becomes worn [43, 75]. The microphones used for it are considered some of the most cost-effective sensor options and are significantly easier to set up, more flexible in their positioning, and less

intrusive on cutting processes than options like dynamometers and acoustic emission sensors [5, 23, 28]. While sound has a longer reaction time than acoustic emission which could lead to a slightly later detection of a tool failure [5], and its sensitivity to signal noise can make it less accurate than other signals initially [23], its features in the time and frequency domains correspond strongly to tool wear and several research groups have confirmed its effectiveness for TCM [1, 7, 22, 23, 28-30, 43, 53].

Achyuth Kothuru, for example, thoroughly compared the strengths and weaknesses of the most common signals used for TCM, and rated sound three out of three stars for cost efficiency, flexibility, non-intrusiveness, and reliability—every category that was assessed except for "accuracy", thereby scoring the highest overall out of the eight signal types evaluated [23]. This one weakness on the accuracy scale presents a valuable opportunity for new advanced analysis techniques to potentially make a significant impact on the current body of TCM research by improving the accuracy that can be achieved using sound. Zhao et al. [43] support Kothuru's evaluation, stating that sound signals are the most appropriate TCM signal for use in production environments due to their low sensor cost and practicality advantages As a wide variety of TCM techniques have proved effective in laboratory environments but no reliable systems have made it to the industrial market [5, 7, 19, 23, 27, 28], a focus should be made on designing future systems for that specific end use. Therefore, the holes which currently exist in the TCM research body with respect to sound signal processing and analysis should be addressed.

2.2.6 Other Indirect Sensors for TCM

A few other sensor types are used less frequently for TCM applications. Process temperatures, for example, raise with the increases in friction caused by progressively dulling tool

edges [9, 12, 18]. This increase in temperature also contributes to the eventual failure of the tool, when high temperatures at the tool tip worsen its mechanical properties [9, 12]. These cutting temperatures can be monitored using tool thermocouples, tool-work thermocouples, radiation techniques, and thermo-chemical reactions [7, 76-78]. Finally, workpiece surface roughness has been found to increase with tool wear, decreasing part quality but also making more data available for TCM [5, 7, 9, 12]. This can be measured using needle-point probe devices, optical devices using a laser or light diffusion, or 3D graphics techniques [5, 7, 12]. However, these methods can be time-consuming and negatively effected by common process conditions such as the use of coolant and the presence of machining chips [7, 43].

2.2.7 Sensor Fusion

As tool wear is a complex phenomenon influenced by many variables, and each process signal type described above has its own advantages and disadvantages, researchers have begun to utilize multiple sensors for TCM through sensor fusion [2, 7, 13, 42, 60, 79, 80]. The aim of this technique is to harness a larger amount of the available process information through the calculation of features from several signals which may complement, reinforce, or introduce new perspectives to each other [2, 13, 79]. In recent years and with the help of faster data processing methods, sensor fusion has shown advantages in the TCM field by reducing uncertainty, reducing sensitivity to sensor noise, and improving algorithm performance [2, 4, 7, 13, 20, 28, 42, 51, 81].

With the goal of developing a TCM model which would be best suited for adoption by the manufacturing industry, the cost effectiveness, intrusiveness, flexibility, and ease of installation of any sensors used must be heavily considered [7]. As production environments have high levels of machining and environmental variability, a sensor fusion technique is selected in order to

maximize the generalization ability of the final models when presented to new unseen conditions [28]. To balance these economic, practicality, and performance considerations, five indirect, cost-effective, and easy to implement process signals were chosen for the following TCM analysis: sound, spindle power, x-axis drive load, y-axis drive load, and z-axis drive load.

2.3 TCM Conventional Analysis Approaches

A wide range of analysis techniques have been applied to the problem of tool condition monitoring over the past several decades, with varying results.

2.3.1 Cutting Parameter Analysis

The relationship between the cutting parameters used for a machining process, such as the cutting speed, chip load, and depth of cut, and the approximate life of a tool has been studied for over 100 years [7, 12, 36, 37]. Researchers have found that while an increase in any of these three parameters will cause an overall decrease in the average tool life across a large number of tool samples, the cutting speed holds the most influence, and the depth of cut holds the least [2, 5, 82-84]. Equations for the cutting speed and chip load in the milling process are given by Equations 2.1 and 2.2, in which N is the spindle speed in RPM, D is the tool diameter, and Z is the number of flutes.

$$cutting\ speed = \frac{\pi * D * N}{1000} \tag{2.1}$$

$$chip\ load = \frac{feed\ rate}{N * Z} \tag{2.2}$$

Several formulas have been proposed to quantify these effects, perhaps the most well-known of which is F. W. Taylor's tool-life formula [36]:

$$\boldsymbol{VT^n = C} \qquad (2.\mathbf{3})$$

where V is the cutting speed, T is the tool life, n is the Taylor exponent, and C is a constant related to the cutting speed which would produce a tool life of one minute. While this estimation provides valuable insight into the effect of the cutting speed on tool life and is helpful for many practical applications, it ignores the significant impact of chip load on tool life, it does not apply to low cutting speeds or to the initial and final stages of a tool's wear, it does not account for variations in the Taylor exponent n, and it requires many experiments in order to determine model constants [2, 26, 37].

Several researchers have since presented alternative tool-life formulas, seeking to better estimate a machining tool's life based on a variety of parameters [37-39]. However, with the high levels of tool life variance that are present even when the same cutting conditions and tooling are used, and the many random variables which can have large effects on a tool's life, tool wear does not lend itself well to a priori tool life prediction [2, 19, 20, 26]. Even very small inconsistencies between tests, such as slight differences in material properties between two workpieces of the same material, inclusion types and locations within a material, and the angle that the tool contacts material edges or irregularities at, can cause significant differences in the immediate and future tool wear rates [2]. For these reasons, immediately accessible information about a tool's current wear level during a machining process could be invaluable for this application [2, 3, 9, 11, 25, 26].

To address this need using only the process' cutting speed, feed rate, and depth of cut, Okokpujie et al. [80] conducted 27 experiments with varying parameters and analyzed the resulting

tool wear patterns using the least squares method [85]. Through this they were able to develop a mathematical model relating each of the three cutting parameters with a tool's wear measurement after a set number of turning passes, which showed itself to be effective at wear measurement prediction under these specific conditions. However, a mathematical model like this would need to be expanded extensively in order to be evaluated for dynamic practical applications. For accurate tool wear qualification to be possible during machining and across the conditions seen in manufacturing centers, it is likely that additional sensor data will be required [4].

2.3.2 Raw Signal Analysis

Many of the first works in the field focused on identifying patterns between signal raw data values and tool wear, and using cutoff values to directly determine levels of wear [7, 62-64]. Using feed and thrust force sensors in 1976, Langhammer showed regimes in which the patterns between the two forces and the tool wear measurement were approximately linear [64]. Soon after, Uehara et al. [60] identified a strong correlation between tool condition and the relationship between feed force and chip load. Colwell compared axial force signals to their corresponding power signals, and found that the ratio of the two was highly dependent on tool wear [86].

2.3.3 Time Domain Analysis

A wide variety of features have been extracted from signals in the time domain to provide greater insight into processes and tool deterioration. Using an acoustic emission signal, as well as the calculation of its root mean square (RMS), skewness, and kurtosis features in the time domain, Kannatey-Asibu and Dornfeld [87] were able to identify significant correlations to tool wear with high levels of sensitivity. Lan and Dornfeld [88] also calculated the RMS of AE signals during

tool wear progression and noticed that at the moment of tool catastrophic failure, a peak RMS signal can be observed in the AE data. Diei and Dornfeld [89] then built upon this research by proposing a formula to relate this peak AE RMS value to the fractured area and the cutting force at failure.

Binsaeid et al. [13] utilized 135 features from the time and frequency domains of their force, vibration, acoustic emission, and spindle power signals, in order to predict three classes of tool flank wear up to a high level of 0.6 mm. These time-domain features included the mean, RMS, variance, skewness, kurtosis, signal power, peak-to-peak amplitude, crest factor, and the burst rate. Other researchers have confirmed the usefulness of these features, as well as others such as the median, standard deviation, maximum, and clearance factor, through related studies [2, 14, 15, 28, 29, 42, 43, 51, 54, 60, 81, 90]. Time-domain features appear to be the most useful for force and force-related signals such as spindle power [12].

2.3.4 Frequency Domain Analysis

For some signal types such as acoustic emission, sound, and vibration signals, analysis in the frequency domain can be the most informative for TCM [7, 12]. To do this, researchers use the Discrete Fourier Transform (DFT) to convert a sampled time-domain signal into a frequency spectrum of amplitudes representing the original signal [2, 5, 12, 13]. For example, Emel et al. [91] found that the AE power within the 400-700 kHz frequency range increased with tool wear progression. Differences were also found between this identified frequency range and that of a tool which had failed catastrophically, and these were used to detect catastrophic tool failure with an accuracy of 84-94%. While Inasaki et al. [92] also identified a gradual increase in AE signal

amplitude with tool wear, they determined the significant peaks to be at frequencies of about 120, 170, and 210 kHz.

Frequency analysis can similarly be a very useful tool for TCM using sound signals, although the relevant frequencies are identified at much lower values than those seen in AE spectra [3, 7, 22, 93, 94]. Through the analysis of turning experiments run on a lathe, Sadat and Raman's study [93] found significant increases in the sound energy contained within frequencies 2.75–3.75 kHz in the initial stages of tool wear. As these peaks were able be minimized by increasing the cutting speed, it was hypothesized that the frequency region's energy increase with tool wear was due to the dulling of the edge and the resultant increase in friction during cutting. L. C. Lee [94] identified a similar characteristic frequency range of 4-6 kHz for a variety of workpiece materials and cutting conditions in turning. Others have found lower frequency ranges, including the milling spindle frequency and its harmonics, to be significant for tool wear [29].

Weller et al. [1] identified a considerable increase in high-frequency vibration energy as turning tools become more worn, and utilized a ratio between the signals' low-frequency (0-4 kHz) and high-frequency (4-8 kHz) energy to detect tool wear. Later, Xiao et al. [95] confirmed this result for the first two stages of wear, but also showed that the ratio then decreased during the severe wear stage. Other features extracted from process signals' frequency domain (FD) which have shown value for TCM applications include the FD mean amplitude, FD frequency center, FD variance, FD standard deviation, FD maximum amplitude, FD RMS, FD skewness, FD kurtosis, FD amplitude sum over a certain frequency range, FD stabilization ratio, and FD crest factor [13, 14, 22, 29, 42- 44].

2.4 Machine Learning Analysis

In the last few decades, machine learning (ML) has proved to be a valuable tool for many areas within manufacturing analytics such as process optimization, fault diagnosis, and TCM [2, 4, 5, 12, 18, 22, 26, 28-30, 42, 43, 45-47, 53, 60, 96-98]. Its advantages in large-scale data processing, autonomous nonlinear pattern recognition, and multivariate analysis have enabled the significant improvement of TCM systems' performance, automatization, and adaptability to new conditions [3, 4, 12, 23, 42]. However, among many ML strategies and model variations studied, a consensus has yet to be reached concerning the optimal configuration of such a ML TCM system [14, 48, 49]. The current research related to several distinct ML strategies will be discussed.

Machine learning models can be split into several descriptive categories. Supervised learning models are generally used for applications in which a label or true value for the ground truth is to be predicted, and the researcher has access to these labels for a training dataset. Using these true labels, the ML model may learn a mapping function which best maps the input data or features to the output labels, and then use this function to predict the labels for new data [23]. Unsupervised learning models, on the other hand, are often used when correct labels for data either don't exist for an application or are not available. In this case, the goal is instead to identify any interesting patterns which relate the samples or features to each other. As a result, this problem is much less defined and its performance is much more difficult to evaluate [23, 99]. Supervised learning is almost exclusively used for TCM applications, as the goals are to accurately predict a distinct metric such as the tool's wear level, wear measurement, or remaining useful life [5, 12, 15, 23, 40, 43, 49, 100]. Research has shown that using calculated features as the input data for a TCM ML model achieves superior results when compared to running a model on the raw signal data in either the time or frequency domain [60, 101].

2.4.1 TCM with Individual ML Models

Several individual supervised learning model types have been evaluated for TCM, including decision tree [49, 50, 96, 102-104], support vector machine [3, 29, 30, 48, 97, 103, 105-107], k-nearest neighbors [29, 103, 108, 109], and artificial neural network [4, 16, 18, 19, 25, 26, 28, 42, 60, 80, 103]. Each of these has shown promising results, along with unique advantages and disadvantages.

2.4.1.1 Decision Tree

The decision tree (DT) algorithm is one of the simplest machine learning models to understand, with a structure similar to that of a flowchart or "tree" [102, 103, 110]. At each node in the tree, the data to be evaluated is split based on a feature condition which is determined during the model's training. The first node, called the root node, is where new data enters the model, and the last level of nodes, called the leaf nodes, gives the predicted classification [102, 103]. As a non-parametric ML model, the DT avoids making any initial assumptions about the data's patterns, and instead is fully open to any patterns which may arise during its training [102].

Decision tree models have been shown to be quite effective for TCM [49, 50, 96, 102-104]. Shurrab et al. [98], for example, compared the performance of six ML model types after they were each trained to identify the binary presence or absence of tool wear in a wax milling study. The decision tree achieved the highest classification accuracy, reported at 99.2%, while the other models to be discussed here, the neural network, k-nearest neighbors, and support vector machine, achieved the 2nd, 3rd, and 4th-best scores at 94.3%, 90.6%, and 87.6% respectively. For the related application of image-based workpiece wear classification, Castejón-Limas et al. [104] found that

out of the DT, k-nearest neighbors, neural network, and Naïve Bayes techniques, the DT model performed best. Elangovan et al. [102] also evaluated the performance of a DT-based TCM model for the turning process, this time with vibration signals. Using simulated wear with measurements of 0.0 mm, 0.3 mm, and 0.6 mm, as well as the unique addition of a "tool tip loose" condition, their model achieved a 10-fold cross validation mean accuracy score of 77.2% between the four tool conditions.

2.4.1.2 Support Vector Machine

The support vector machine (SVM) [111, 112] is also effective for TCM applications, and has been more frequently studied in this context than the DT [3, 29, 30, 48, 57, 97, 103, 105-107, 113]. It takes a different approach to classification, by attempting to identify the optimal set of hyperplanes and support vectors necessary to split up data in a feature space into regions based on the classification labels [97, 103]. After training, these regions determine the label predictions for input test data. For non-linear data such as that which is related to tool wear, a kernel function can be selected to transform the input data into a higher-dimensional space and improve the model's classification performance [43, 103]. SVMs have been found to have an advantage over other ML techniques such as neural networks when classifying and generalizing small-scale samples [14].

The value of SVM techniques for TCM has been proven repeatedly. For example, as part of their force-based TCM ML study, Wang et al. [48] found that their SVM model using a Gaussian kernel function performed better than both the hidden Markov model (HMM) and the radius basis function network, showing a 4-level wear classification accuracy of 93.9% when tested on experiment data from using the same cutting parameters. Kothuru et al. [3] chose an SVM model as the decision-making algorithm for their study on multiple microphone sound signal TCM. They

identified quite interesting results, with the combination of multiple microphone signals improving the SVM classification accuracy from about 89.1% for one microphone, to 97.0% for 3 microphones for 6 wear levels up to 0.15 mm flank wear. However, it is likely that this addition of more microphones within the machine would significantly reduce the practicality benefits of using a sound signal over force and other more invasive signal options. Zhou et al. [43] recently proposed a novel SVM-based strategy to overcome the high levels of signal noise often found in milling process sound signals. By only utilizing a small number of calculated signal features and a unique 2-layer network structure, they were able to improve the model's learning ability for the complex nonlinear relationships between tool wear and sound signal features, and improve TCM performance.

2.4.1.3 K-Nearest Neighbors

While k-nearest neighbor (kNN) strategies have been less thoroughly studied for TCM than SVMs, they have also shown promising results [4, 16, 18, 19, 25, 26, 28, 42, 60, 80, 103]. Their classification process works simply by assessing the location of each testing data point in a feature space and comparing it to the locations of previous training data points [103, 108]. The user-prescribed parameter 'k' then determines the number of closest training data points which contribute to a majority voting process for the final class prediction decision. As the kNN technique does not require a model to be built during the training phase, computation time can be reduced [103, 108].

Li et al. [28] built upon the research by Kothuru et al. [3], described previously, by applying the blind source separation technique to identify a single source signal from the 3 microphone signals, as well as by comparing the performance of kNN, SVM, random forest, and DT

classification models for TCM. Using a training set size of 50%, the SVM and kNN achieved the highest overall classification accuracies at 98.5% and 97.8% respectively. The optimal k-value for this dataset was determined to be k=15. Jegorowa et al. [108] applied k-nearest neighbors to monitor the condition of drill bits between 3 wear levels, while cutting particleboard. Several signals were collected including feed force, cutting torque, vibration, sound, and acoustic emission. Using a selected value of 12, the kNN model achieved an overall accuracy of 76%.

2.4.1.4 Artificial Neural Network

Artificial neural networks (ANN) are a commonly studied ML technique used for TCM [4, 16, 18, 19, 25, 26, 28, 42, 52, 60, 80, 103]. These networks are designed based on the human brain, and attempt to identify patterns within complex data in a similar manner [5, 103]. They are made up of interconnected layers of nodes, with an initial input layer representing the input features, a final output layer representing the number of final classes, and a number of hidden layers between the two which use non-linear activation functions for decision-making [4, 5, 103]. Model performance varies greatly depending on the number of hidden layers selected, as well as the numbers of neurons designated for each hidden layer [4, 5]. Although many variations of ANNs exist, Sick et al. [4] found that about 70% of the neural network research in TCM studied a feedforward paradigm called the multilayer perceptron (MLP), designed by Frank Rosenblatt in 1958 [114]. While neural networks provide high levels of fault tolerance, adaptability, and noise suppression capability [115], they can appear to be underfitting or overfitting if the sample size is insufficient or if the sample noise is high [14].

Artificial neural networks have been applied to TCM in various ways. Mannan et al. [53], for example, trained a neural network to classify tool condition between 3 wide flank wear levels during steel turning based on both images of machined surface texture and sound signals.

Hsieh et. al. [60] used a neural network and accelerometers to analyze features extracted from mico-milling vibration signals' time and frequency domains, distinguish "worn" from "unworn" tool signals, and study the effect of signal bandwidth on classification results. Ghosh et al. [28] used a neural network, along with sensor fusion of force, vibration, spindle current, and sound signals, to predict tool wear measurements during the milling process. They also assessed correlation patterns between the different process signal types and found a high level of correlation between the spindle power and cutting force signals, as well as a "fair" correlation between the sound and cutting force signals. By analyzing the effect of sensor fusion for multiple sensor groupings, it was found that while combining the two axial forces, the spindle motor current, and the spindle motor voltage provided the best results, a combination of only two signals, the spindle motor current and the sound pressure level, also performed very well. As a significantly more practical option for manufacturing centers than the use of expensive force dynamometers, these results were promising.

2.4.1.5 Other Individual ML Models

A few other machine learning model types have been applied to address the tool condition monitoring problem, as well. Using force signals, Wang et al. [40] studied how the Hidden Markov Model, developed for speech recognition and now spreading to manufacturing insight applications [6], could be used to predict the state of tool wear. Kannatey-Asibu et al. [51], Wang et al. [106], and Wang et al. [116] also showed that Hidden Markov Models can be effective for TCM.

However, Wang et al. [48] showed that both SVM and ANN models performed better than a Hidden Markov Model for TCM.

Regression analysis has also been investigated, with mixed results. While one study [40] concluded that Gaussian Regression had potential for remaining tool life prediction in milling based on a small number of samples, and Korkut et al. [100] found that both regression analysis and ANN techniques performed adequately for predicting the temperature at a tool-chip interface, the linear regression model investigated by Shurrab et al. [103] performed the worst out of six ML model types compared for tool wear level classification. By distinguishing between local and global linear regression analysis, however, Wu et al. [15] found that the Saucer's local linear model showed better results than both the global linear model, and a neural network. While regression can be used to gain a higher level of resolution on a tool's wear, and updating a model's information mid-process can help it achieve strong results, the re-training required for this technique makes it less practical for TCM than classification models due to in-process computation time limitations [16].

2.4.2 TCM with Ensemble Machine Learning

Ensemble machine learning methods are a technique in which multiple base learners are combined to achieve superior performance by leveraging "the wisdom of the crowd" [49, 117]. They utilize the collective knowledge and insights generated by the multiple base learners, and therefore can benefit from each base algorithm's individual advantages in different situations [118]. As a result, and across many applications, ensemble models generally outperform individual methods by improving model robustness and accuracy, as well as reducing variance [49, 117-120]. The individual learners that make up the ensemble may use completely different decision models,

or they may be identical to each other but trained on different subsets of data [49, 119]. Ensemble models using these distinct strategies are known as heterogeneous or homogeneous ensembles, respectively.

In recent years and with technological advancements to decrease computation time, a few researchers have begun to assess the effectiveness of ensemble machine learning techniques for TCM [6, 13, 14, 48-52]. However, the current research is limited and there remains much to be uncovered.

2.4.2.1 Heterogeneous Ensemble Machine Learning

Heterogeneous ensemble ML is a type of ensemble machine learning strategy in which multiple unique base models are combined. It is generally beneficial for these base models to be substantially different from each other, so that the ensemble model can draw from as many insights and individual advantages as possible [13, 50]. The main groups of heterogeneous models include voting schemes and stacking classifiers.

2.4.2.1.1 Hard Voting

Ensemble Hard Voting (EHV), sometimes called "Majority Voting," is one of the most commonly studied heterogeneous ML techniques due to its simplicity, robustness, and stability [13, 40, 51, 121, 122]. In this strategy, each of the unique base models is trained on the same training dataset, and each base model's classification output is used to determine the ensemble model's final classification. For Hard Voting, each model gets one "vote" for its output classification, and the class with the largest number of votes will be the ensemble model's final classification [13, 123].

Using HMM, Bayesian Rule, Gaussian Mixture, and K-Means base models, Kannatey-Asibu et al. [51] studied how Hard Voting might be applied to TCM and found that the ensemble model's performance could be improved by configuring multiple types of weightings based on the individual models' classification performance. Acoustic emission signals were used for this research. Wang et al. [48] utilized a Hard Voting ensemble model and force signals to classify tool wear into 4 levels based on SVM, HMM, and radial basis function (RBF) neural network base models. Based on their data from one tool run to failure, the Hard Voting ensemble was found to perform better than any of the individual base models, but slightly less well than the stacking ensemble model also studied. With SVM, ANN, and RBF neural network base models, Cho et al. [122] and Binsaeid et al. [13] also confirmed the improvement in TCM classification performance when Hard Voting is used compared to when the individual ML models are used. However, with only these studies into the application of Hard Voting for TCM having been reported currently, there remain several questions about the optimal base model selections and the generalization ability of TCM Hard Voting models to new environmental or machining conditions.

2.4.2.1.2 Soft Voting

For the ensemble Soft Voting (ESV) strategy, however, it is not the base models' output classifications, but rather each one's estimations of individual class probabilities, which are averaged and considered by the final ensemble classifier [124-126]. This of course relies on each base model calculating a probability metric which can be used, which is something that not all model types do [124]. When applicable, this Soft Voting strategy is considered to have better flexibility and generalization than Hard Voting, and is recommended over Hard Voting for well-calibrated base models [121, 123].

While Soft Voting has been successful in various other fields such as medical diagnosis [127], online emotion classification [128], semiconductor manufacturing defect detection [129], and bearing fault detection [126], it has not yet been studied within TCM. In the semiconductor manufacturing defect study, Saqlain et al. [129] found that a Soft Voting ensemble classifier performed better than not only the base models, but also better than the Hard Voting ensemble classifier. In perhaps the most closely related study, Li et al. [126] used vibration signals to classify between four distinct bearing fault types, and assessed the impact of Soft Voting on fault classification accuracy and stability. Across signals augmented with eight levels of additive white Gaussian noise (AWGN) to simulate a variety of environmental conditions, the use of Soft Voting dramatically improved the bearing fault classification accuracy for all 8 conditions. In addition, it improved the prediction stability for all 5 noise levels with sound-to-noise (SNR) ratios above 0, but not the 3 which were at or below 0, suggesting that while Soft Voting generally improves prediction stability, "when there is too much noise, the [base classifiers are] not certain about [their predictions] and aggregating multiple predictions does not help to increase this certainty" [126]. Based on its success in various fields, including an application which is closely related to TCM, the Soft Voting ensemble method is a promising technique for TCM, and a significant research gap.

2.4.2.1.3 Stacked Generalization

The stacked generalization, or "stacking" method introduced in 1992 by David H. Wolpert [130], is a different type of heterogeneous ensemble ML method in which the output classifications given by the base models are then fed into a higher level ML model, or "meta-learner," as effective training or testing features [13, 50]. With this method, the meta-learner has the opportunity to

identify significant patterns between the outputs of different base learners, and potentially improve the overall classification accuracy [50]. While stacking meta-learners are typically based on SVM, logistic regression, or random forest techniques [120], only SVM-based stacking has been applied to TCM.

This type of heterogeneous ensemble ML has only begun to be studied for the TCM application. In 2014, Wang et al. [43] created a stacked ensemble model based on SVM, HMM, and RBF base models, as well as an SVM meta-learner. They then compared its results for tool wear classification to those of a Hard Voting ensemble model and three homogeneous ensemble models based on SVMs, HMMs, and RBFs, to find that the stacked ensemble model achieved the best average accuracy and classification stability among all five ensemble models and three base models studied. However, these results were based on data from only one experiment, and therefore lack generalization capability to machining or environmental condition variations. Hui et al. [50] then compared the performance of a stacked ensemble model to Bagging-SVM and AdaBoost-SVM homogeneous ensemble models, as well as ANN and SVM individual models. Using SVM, DT, and naïve Bayes base models, as well as an SVM meta-learner, they found that for their one experiment, the stacked SVM model achieved the best accuracy and stability when trained on 80% of the experiment's data and tested on the other 20%.

Finally, Binsaeid et al. [13] ran eight milling experiments with different depth of cut, cutting speed, and chip load machining parameters to evaluate the prediction of a tool's wear level between three levels from 0 to 0.6 mm flank wear. A stacked ensemble model using an SVM meta-learner and SVM, ANN, and RBF base learners was shown to perform better than a Hard Voting ensemble technique as well as the individual models. 10x-repeated 10-fold cross validation was used for performance evaluation, and four sensor types were also evaluated with force achieving

the best results. Due to the machining parameter variation study and the 10-fold CV performance metric, these results give a good idea of different models' performance across a range of machining conditions and allows the model techniques to be compared. However, they do not provide insight into the models' performance for individual machining parameter values or specific parameter changes. As a result, the transferability of a stacking ensemble TCM model to specific cutting parameter variations still has yet to be evaluated.

2.4.2.2 Homogeneous Ensemble Machine Learning

Homogeneous ensemble machine learning is a type of ensemble ML technique in which multiple base models of identical structures are combined. These base learners' initiation parameters, weight values, or training data subsets may be very different, however [48, 49]. This allows several variations of the base algorithm to be run, and their results to be combined, to generally achieve superior results than through the use of just one base model [48, 49].

2.4.2.2.1 Random Forest

Along with Hard Voting, random forest (RF) is another commonly studied ensemble ML method in other industries, and it has seen a few previous applications to TCM [14, 29, 49, 54, 113]. It is a homogeneous ensemble method, created in 2001 by Leo Breiman [131] and made up of several decision trees trained on different randomized subsets of the overall training dataset using sampling with replacement. This process is called bootstrap aggregation, or "bagging" and helps reduce overfitting [54]. As decision trees are very sensitive to their specific training dataset, this process results in many decision trees of different configurations. The diversity of the trees' designs is also further randomized by limiting each DT node's possible feature choices to a random

subset. The random forest then takes the mode of the individual trees' class predictions as its final prediction [49]. It has become a well-known ensemble ML method due to its high accuracy, its efficiency for processing large datasets and feature numbers, its tendency to avoid overfitting, its tolerance to outliers and noise, its high generalization ability, and its effectiveness for unbalanced class sizes [14, 54].

The random forest algorithm has been shown to be effective for TCM in a limited range of studies. Riego et al. [49], for example, used computer vision and multiple ML algorithms including a random forest to assess a milled holes' surface texture and classify a tool as "worn" or "unworn". Wu et al. [54] applied a random forest to tool wear progression analysis using dynamometer, accelerometer, and acoustic emission data from one tool life experiment and one set of experimental conditions. They found that for a random forest of 10,000 trees, sufficiently accurate tool wear measurement predictions could be obtained. Yuan et al. [14] applied a RF to TCM using a spindle current signal and a larger set of machining conditions. They found that for a large wear range of over 0.5 mm, a random forest could be used to effectively classify the wear into three levels, as well as perform better than both an SVM and an ANN. However, as the models were always trained and tested on data from the same experiment each time, the transferability of the RF to new specific machining conditions or individual tools was not assessed, and this remains a research gap.

2.4.2.2.2 Extremely Randomized Trees

Extremely Randomized Trees, or "Extra Trees" (ET), an algorithm proposed by Pierre Geurts in 2006 [132], builds off of the random forest design, with the adaptation that instead of letting each decision tree node identify the optimal feature threshold for splitting, these thresholds

are now assigned random values [49]. This increases the randomness of the individual DT formations, improves the overall model's ability to avoid overfitting, increases the generalization ability, and reduces model computational complexity [49, 133].

The Extra-Trees algorithm has only been applied to TCM by one research group previously: Riego et al. [49], in 2020, included it in their study of nine machine learning algorithms, including a RF, three ensemble boosting algorithms, a DT, a SVM, and a "dumb classifier," for surface quality assessment and indirect TCM. Based on workpiece hole surface images and with the aim of automatically identifying substandard workpiece surface quality caused by tool wear, their team found that the Extra-Trees model achieved the highest accuracy and F1 scores out of all nine ML models implemented. The results of this study, along with the successes of the similar RF algorithm, showcase the potential for the ET model to be highly effective for TCM. Furthermore, the effectiveness of ET models for similar manufacturing monitoring tasks such as wireless sensor network fault diagnosis [134], wind turbine generator fault diagnosis [133, 135], and additive manufacturing temperature profile prediction [136], which many times surpasses the effectiveness of individual models [134, 135] and other ensemble models [133, 134], further supports this point. With Extra-Trees having only ever been applied to TCM indirectly using surface image data, there remain many opportunities for increased study.

2.4.3 Model Generalizability

For industrial manufacturing processes, including the especially adaptable milling process, machining conditions are frequently changed in order to suit the shapes, materials, and part quality required of a finished product. In addition, a shop's environmental conditions such as background noise also change significantly based on other machines being run, the time of day, and many

effectively random environmental variations. The previously proposed TCM solutions' lack of adequate generalizability to these various machining and environmental conditions is a main reason for their lack of adoption by industry manufacturers [19, 22, 23, 28-30, 32]. For these reasons, it is important to understand how a trained TCM system's performance may change when the system is transferred to new unseen conditions.

2.4.3.1 Study of Model Generalizability in TCM

While it is agreed upon that the previously developed TCM solutions generally perform significantly worse in real-world production settings than they do in laboratory experiments, and that variations in cutting and environmental conditions play a role in this, the specifics of this effect are not well understood [19, 22, 23, 28, 29, 30, 32]. Many studies in the field of TCM collect data from only one experiment, and draw conclusions based on the performance of ML models that are trained, tested, and validated all on data from one individual tool's life [32]. These results do not account for the significant variability in wear rates and patterns that occur even under the same machining settings and therefore lack generalizability to other experiments [32, 19].

In order to improve the systems' generalizability to new machining conditions, several research groups have also trained TCM models on signal data from tool wear experiments using a variety of different cutting parameters [3, 6, 9, 13-15, 19, 22, 43, 74, 94]. However, this data is often then aggregated and split into training and testing sets without regard to which data came from which parameter sets or experiments. While this makes the models more generalizable to new conditions than if only one experiment's data were used, it does not allow the effects of individual condition changes on model performance to be studied. In addition, previous research has very rarely reported data for repetitions of parameter sets [2], and as a result there is little

understanding of how ML TCM models perform on data from new experiments which were conducted using the same parameters.

Several other strategies for increasing TCM model generalizability have also been investigated, including through sensor fusion as discussed in Section 2.2.7, by filtering out high or low frequency signal components to reduce noise [2, 137], and by using multiple of the same sensor type in different locations to better identify the source signal [3, 22, 23, 29]. As discussed previously, the combination of signals from multiple sensor types, called sensor fusion, has been shown by several research groups to improve model generalizability and reduce their sensitivity to noise [2, 4, 7, 13, 20, 28, 42, 51, 81]. The use of a low-pass filter to remove high-frequency signal components in an attempt to reduce noise, however, is not as optimal. According to Larry P. Heck, "While relatively easy to implement, these techniques have proven to be generally ineffective at reducing the noise and tend to remove information necessary for proper tool wear classification" [137]. Finally, the use of multiple sensors of the same type has also been investigated for TCM. Kothuru et al. [3], for example, used three microphones located in different positions relative to the milling process to study how well a SVM-based TCM model trained on data from certain individual microphones was able to transfer to signals from a new microphone. It was found that training the model on data from two microphones' data always performed better when tested on the third microphone, than when the model was trained on data from only one microphone. This showed that increasing the number of sensors used can improve model generalizability even when they are of the same type. Li et al. [29] then built on this with the aim of reducing noise in a signal through blind source separation, and found the technique capable of improving TCM performance. While several methods have been presented for noise removal from process signals for TCM, with mixed results [2, 29, 30, 137-139], there has been little study of the

effect of noise addition, and no analysis of if it can be used to improve TCM model generalizability.

2.4.3.2 Noise-Based Dataset Augmentation

In various other research areas, additive noise has begun to be studied as a method of data augmentation for the improvement of machine learning generalization to data which is different from the original training data.

For neural networks, especially, several studies have shown that the addition of noise, or "noise injection," to training dataset signals or features can reduce model overfitting and improve generalization ability to new data [34, 35, 140-144]. According to Bishop, "heuristically, we might expect that the noise will 'smear out' each data point and make it difficult for the network to fit individual data points precisely, and hence will reduce over-fitting. In practice, it has been demonstrated that training with noise can indeed lead to improvements in network generalization" [140]. Specifically, it has been concluded that ANN training with noise is equivalent to a form of regularization with an extra term in the error function [34, 144]. Guozhong An [35] studied how noise added to the inputs, outputs, and weights of an artificial neural network during backpropogation training affected its objective functions and testing outputs. It was found that while output noise had no effect on the network's generalization ability, both input noise and weight noise smoothed the ANN output functions, weight noise improved model generalization for the classification problem, and input noise improved generalization for both the classification and regression problems. Audhkhasi, Kosco, and Osoba have also showed that additive noise can speed up algorithm convergence during the training of multilayer ANNs [145, 146].

While training ANNs on datasets with additive noise, also known as the "noisy training" approach, has been a known tool in the neural computing community for decades, it has only recently begun to be studied for direct industrial applications [33-35]. In 2015, for example, Yin et al. [144] investigated multiple types of additive noise and their relation to deep neural network speech recognition performance. They found that white noise injection into training datasets not only dramatically improved the speech recognition results when the model was tested on other white noise-augmented data; it also successfully improved model generalizability to various other noise conditions when the level of white noise used for training was low. The use of a combination of no-noise data, injected white noise, and either cafeteria or car noise for training set injection also significantly improved the model performance for many other noise conditions. The use of combinations of sound levels in the training dataset was also shown to improve model generalizability. These results were a significant step forward in speech recognition model development, and highlighted a promising opportunity for model generalizability analysis in other areas such as tool condition monitoring.

This noisy training approach has been studied for other ML model types to a much lesser extent than for ANNs. In 2021, Xing et al. [145] showed that through asymptotic regret analysis, noise injection into training datasets for a K-Nearest Neighbors model does not significantly change its predictive performance on new data, as this algorithm is already relatively robust to small variations in its samples. In fields outside of TCM, some of the ML algorithms addressed in this study, DT, SVM, kNN, EHV, and RF, have been evaluated in terms of their robustness to noise in either the testing or training datasets, when the training and testing groups were sampled from the same dataset [147-151]. In all of these cases, however, the models' generalization ability was not fully measured as there was no significant difference between the original training and

testing sets, and as a result the addition of noise to the training datasets generally had a negative effect on model performance [148, 149]. No studies have been found in any field which apply the noisy training approach of injecting noise into a training dataset with the aim of improving a model's generalizability to significantly different data using DT, SVM, EHV, ESV, Stacked SVM, RF, or ET models [152]. In addition, this noisy training approach has never been studied in the field of TCM using any ML model type.

2.5 Summary of Research Gaps

Several gaps currently remain in the research surrounding tool condition monitoring, and contribute to the industry's lack of access to a reliable, economical, and adaptable solution [19, 22, 23, 28-30]. Ensemble machine learning is a promising technological advancement which has shown substantial robustness, stability, and accuracy advantages for classification problems in related fields, but has only been applied to a limited extent within TCM [31, 108-111]. Specifically, Soft Voting has never been applied to TCM, and Extra-Trees has never been applied using process signals of any kind, despite both models showing significant reliability and generalization advantages over individual models and previously studied ensemble methods. In addition, it is not well understood how the various ML techniques and configurations, including individual, heterogeneous ensemble, and homogeneous ensemble methods, compare for tool wear classification.

In addition, the sound signal, which has been identified as one of the best process signals for practical application due to its low cost, ease of installation, and lack of process interference [5, 23, 28], has never been studied using ensemble ML for TCM. While its lower accuracy compared to other process signals has limited its application before [23], it is not currently

understood how more reliable algorithms such as ensemble ML techniques, in combination with the advantages of sensor fusion, may affect this metric.

Previous systems' inability to adapt to the highly variable machining and environmental conditions present on shop floors has also been a great hindrance to their industrial adoption, and these generalizability effects are still not well understood [19, 22, 23, 28-30, 32]. While a few ensemble ML-based TCM algorithms have been trained using data from multiple cutting conditions, the their generalizability across specific cutting parameter changes has not been evaluated or compared [6, 13, 14]. In addition, the noise injection method of improving ML model generalizability to new environmental conditions by training on noise-augmented data, has never been studied for TCM. Each of these research gaps, if addressed, would advance the current TCM knowledge base and further enable a practical TCM solution.

CHAPTER 3: METHODS

In order to evaluate the research questions and fill these research gaps, a milling tool wear experiment was designed in order to collect a diverse original dataset to be used for the TCM model evaluation.

3.1 Experimental Design

The experiments for this study were run using an EMCOMILL E350 3-axis vertical CNC machine with a maximum spindle speed of 10,000 RPM, which is equipped with a Siemens Sinumerik 828D controller. A combination of internal and external sensors were used for the analysis. The spindle power and the x, y, and z-dimension axial load signals could be extracted directly from the Siemens controller, as these are automatically monitored for the machine's real-time dynamic control. Using the controller's Trace function, these signals were recorded at a sampling frequency of 166 Hz and extracted at specific times during each experiment. While the spindle power was extracted directly in units of W, and the axial loads were initially reported in terms of % of the machine's maximum axial loads. Given the machine's maximum loads of 3000, 3000, and 4000 N in the x, y, and z directions respectively, these values could be converted to direct force measurements [20, 153].

Two external sensors, a PCB Piezotronics model 130F20 ICP electret array microphone and a Dino-Lite AF3113T microscope were also used. The microphone was attached 2.5 feet away from the machining process, in the top-left region of the machine and far from any potential machining processes. The robust microphone design, the easy and secure installation, and the non-intrusive location were selected for their high practicality for industrial applications. The

microphone was sampled at a frequency of 44.1 kHz using a Model 485B39 Digital ICP USB signal conditioner and the MATLAB Audio Labeler application. In addition, an analog sound calibrator was used to check the accuracy of the collected sound signals. Given the microphone's sensitivity of 45 mV / Pa, the output voltage values could be converted to sound pressure values in Pascals.

The Dino-Lite AF3113T microscope was used for tool wear imaging after each set of three machining passes. With a resolution of 640 x 480 pixels and a magnification of 20x – 50x, it was able to capture clear and measurable images of the flank wear's progression. A calibration scale was also used to confirm image dimensions during the wear measurement. At the magnification and positioning used, an image resolution of 44.5 pixels per mm was obtained which, especially considering the grouping of wear into distinct classification groups, is appropriate for this application. An enclosure was used to protect the microscope from the coolant used during cutting.

The machine enclosure was set up as shown in Figure 3.1 for the experimentation.

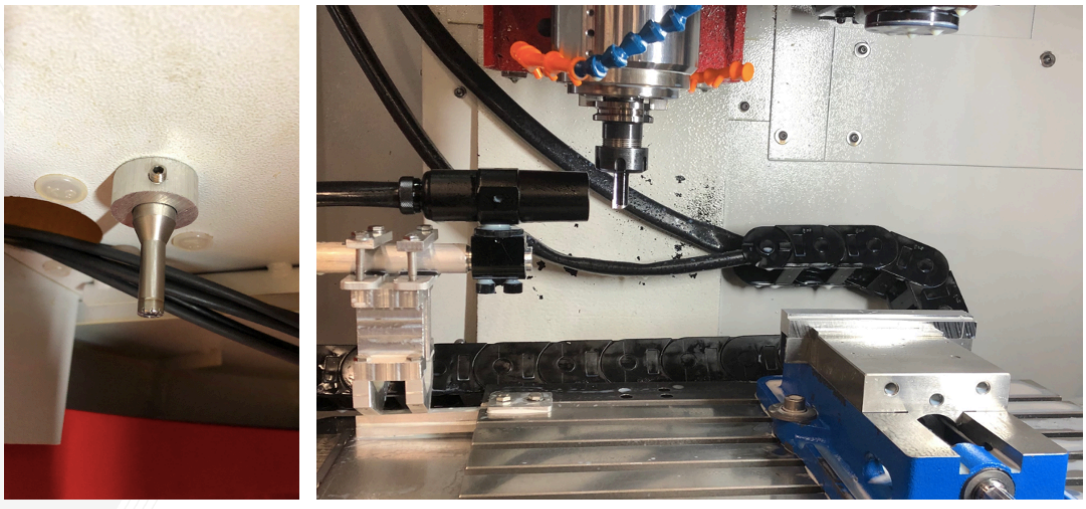


Figure 3.1: Machine layout for experimentation. The (a) microphone position and (b) tool position for imaging are shown

A 6 x 1 x 2 inch (15.24 x 2.54 x 5.08 cm) block of annealed D2 tool steel was used as the workpiece. This high-carbon, high-chromium tool steel is used for applications such as dies, punches, forming rolls, and shear blades, and has a Rockwell hardness rating of B95 in the annealed condition. This material was chosen in order to achieve reasonably short tool lives for all experiments run, and the annealed condition was chosen to match the state which is most commonly used for machining D2 in industrial applications. Coolant was also used in order to match real-world conditions, to target regular abrasive wear instead of thermal cracking, and to ensure that the final ML models are robust to signal noise caused by its use.

End milling passes were run across the workpiece along its longest dimension in the x direction, and the tool was positioned for wear imaging after each group of three machining passes. The tool was wiped down before each image to remove coolant and chips which could interfere with wear measurement. The tooling used for the experiments were 0.5-inch diameter 1-flute Kennametal KICR end mill bodies equipped with square-shaped TiAlN-PVD-coated carbide Kennametal KIPR inserts designed for milling steel.

Eight experiments were run, under the four machining conditions shown in Tables 3.1 and 3.2. In order to assess how the TCM models performed under specific parameter changes, both the spindle speed and feed rate were varied between high and low conditions. All four of these resulting parameter sets were within the manufacturer's recommended range of operating conditions. The axial depth of cut, radial depth of cut, and coolant use were kept constant for all tests. Each parameter set is repeated twice in order to better account for and study tool wear's inherent variability when machining conditions are unchanged.

Table 3.1: Experimental overview

Experiment #	Parameter Set	Spindle Speed	Feed Rate
1	A	HI	HI
2	B	HI	LO
3	C	LO	HI
4	D	LO	LO
5	D	LO	LO
6	C	LO	HI
7	B	HI	LO
8	A	HI	HI

Table 3.2: Experiment cutting conditions

Parameter Set Label:	A	C	B	D
Variable Setting Labels:	HI, HI	LO, HI	HI, LO	LO, LO
Spindle Speed (rpm):	4400	3700	4400	3700
Feed Rate (in/min):	34	34	24	24
End Mill Diameter (in):	0.5	0.5	0.5	0.5
Number of Inserts per Tool:	1	1	1	1
Axial Depth of Cut (in):	0.15	0.15	0.15	0.15
Radial Depth of Cut (in):	0.025	0.025	0.025	0.025
Coolant (on/off):	on	on	on	on
Cutting Speed (ft/min):	576.0	484.3	576.0	484.3
Chip Load (in):	0.0077	0.0092	0.0055	0.0065
Chip Load (mm):	0.196	0.233	0.139	0.165
Material Removal Rate (in^3/min):	0.128	0.128	0.090	0.090
Material Removal Rate (mm^3/min):	2089	2089	1475	1475

Each experiment is continued until a maximum flank wear of 0.3 mm is reached, in accordance with the international standard for tool life testing in milling, ISO8688-2 [55], as well as several previous works [16, 26, 68, 154, 155]. Although some studies have found the third stage of wear, severe wear, to start at flank wear measurements above 0.3 mm under certain conditions [28, 43, 57], the use of a tool past approximately 0.3 mm will begin to cause inadequate dimensioning and low part quality for most applications, no matter which wear stage it is at, so the tool should be changed by this level [55, 68, 154, 155]. In addition, a TCM model which could achieve high classification accuracy without relying on the signal changes from significantly higher wear levels than those at which a tool should be used, or a third wear stage which may occur

at varying wear measurements, would be especially beneficial. For these reasons, the 0 – 0.3 mm wear range is chosen as the focus for this study.

The data will then be labeled based on which range of wear values its measurement falls into. For a wear resolution of three levels, the data will be split according to Table 3.3.

Table 3.3: Tool wear ranges for 3 classes

Wear Level	VB Range (mm)
1	0.0-0.1
2	0.1-0.2
3	0.2-0.3

These wear levels will be the classes which the TCM ML models will aim to predict.

3.2 Data Pre-Processing

The data collected from the eight experiments were then processed on a laptop computer using the Python IDE PyCharm to become usable as training and testing data for the machine learning algorithms. To do this, first the relevant machining data for each signal type must be separated from the microphone and controller process signals in which machining was not being done. In addition, the first one and the last two passes of each layer of the workpiece are also excluded, as they showed consistent variations in the process signals which were unrelated to tool wear. Similarly, the tool's entry and exit times from the workpiece are also excluded, as during those times the tool is not consistently engaged with the workpiece and process mechanics are changed. The time required for the tool to travel one radius in or out of the material is calculated for each of the two feed rates used, and this time is excluded from both sides of each included machining pass signal. From the remaining machining time for each pass, 9 seconds of data are taken for experiments with a feed rate of 34 inches per minute, and 13 seconds of data are taken

for a feed rate of 24 IPM. Then, the corresponding one-second segments from each of the five signal types are linked to form one ML sample. This data segmentation method produces a large number of samples to be assessed independently for the machine learning analyses, and is reasonable due to the machining signals' relatively time-invariant nature when evaluated on the order of seconds under the consistent operating conditions described [2, 22, 23, 29, 30, 48].

As the tool's wear is only measured after each 3 passes (either 27 or 39 1-second ML samples, depending on the feed rate used) and abrasive flank wear is considered a gradual process, the wear was assumed to increase linearly between each pair of wear measurements. Therefore, the wear level cutoffs may be crossed at any time and is not held constant between images.

79 features are then extracted from the signals of each one-second ML sample, based on features found to be highly correlated to tool wear or related conditions by previous works [2, 13-15, 22, 28, 29, 42-44, 51, 60, 81, 90, 93, 94]. As features from both the time domain (TD) and the frequency domain (FD) have been shown previously to be strong predictors for tool wear, both analysis techniques were employed for this study. The full list of extracted features is displayed in Table 3.4, with 19 features being extracted from the sound signal and 15 being extracted from each of the four controller signals. The relevant feature formulas are shown in Table 3.5. As the sound signal is symmetric about a value of 0 Pa, the absolute value was taken of each sound pressure reading prior to time-domain feature extraction.

Table 3.4: Extracted features

Sound Features:	Controller Signal Features (spindle power & axial loads):
TD mean	TD mean
TD standard deviation	TD median
TD kurtosis	TD standard deviation
TD skewness	TD kurtosis
TD maximum	TD skewness
TD median	TD maximum
TD RMS	TD RMS
TD range	TD range
TD crest factor	TD crest factor
TD clearance factor	TD clearance factor
FD maximum amplitude	FD maximum amplitude
FD peak at spindle frequency	FD peak at spindle frequency
FD peak at 2x spindle frequency	FD amplitude sum 0-83Hz
FD amplitude sum 0-500Hz	FD frequency center
FD amplitude sum 0-9000Hz	FD frequency standard deviation
FD amplitude sum 4-6kHz	
FD amplitude sum 2.75-3.75kHz	
FD frequency center	
FD frequency standard deviation	

Table 3.5: Feature formulas

Time-domain features:

(x_i is a signal data series for i = 1, 2, 3, ..., N, where N is the number of data points in one segment.)

Mean $x_m = \frac{\sum_{i=1}^{N} x_i}{N}$

Standard deviation $x_{std} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - x_m)^2}{N-1}}$

Kurtosis $x_k = \frac{\sum_{i=1}^{N}(x_i - x_m)^4}{N * x_{std}{}^4}$

Skewness $x_{sk} = \frac{\sum_{i=1}^{N}(x_i - x_m)^3}{(N-1)x_{std}^3}$

Maximum $x_{max} = \max|x|$

RMS $x_{rms} = \sqrt{\frac{\sum_{i=1}^{N} x_i{}^2}{N}}$

Crest factor $x_{crest} = \frac{x_{max}}{x_{rms}}$

Clearance factor $x_{clear} = \frac{x_{max}}{(\frac{1}{N} * \sum_{i=1}^{N}\sqrt{|x_i|})^2}$

Frequency-domain features:

(s_k is a DFT frequency spectrum for k = 1, 2, 3, ..., K, where K is the number of spectrum points. f_k is the frequency at the kth spectrum point.)

Frequency mean $s_m = \frac{\sum_{k=1}^{K} s_k}{K}$

Frequency peak $s_{max} = \max(s_k)$

Peak at spindle frequency $s_{spindle} = \max(s_k)$ for f within 5% of the spindle frequency

Frequency sum $s_{sum} = \sum_{k=1}^{K} s_k$ within a specified range of f

Frequency center $s_{fc} = \frac{\sum_{k=1}^{K} f_k s_k}{\sum_{k=1}^{K} s_k}$

Frequency standard deviation

$s_{std} = \sqrt{\frac{\sum_{k=1}^{K}(f_k - s_{fc})^2 * s_k}{\sum_{k=1}^{K} s_k}}$

For the time domain feature calculation, all data points in a signal's 1-second segment are used to calculate a feature directly. In order to calculate the frequency domain features, however, a Discrete Fourier Transform must first be applied to convert the data from a time domain signal to a frequency domain amplitude spectrum showing the various frequency components which make up the time domain source signal. From there, the frequency domain spectrum can be used to calculate the relevant frequency-based features. As a frequency component can only be accurately sampled if the frequency is below the Nyquist frequency, or half the sampling frequency, this frequency spectrum only spans up to half the sampling frequency. As a result, the sound signal's frequency spectrum can be assessed up to about 9 kHz due to the signal conditioner's maximum sampling frequency of 20.7 kHz +/- 5%. The spindle power and axial load signals, sampled at 166 Hz, can be assessed up to a frequency of 83 Hz, which is high enough to measure and track the spindle frequency peak with respect to tool wear.

After all 79 features are calculated, their distributions are standardized. This is an important preprocessing step to ensure that the large differences in original feature ranges and mean values do not allow some features to overpower others in distance-based algorithms such as kNN and SVM. This step has also been shown to significantly shorten the training convergence time for ANNs [156]. Finally, with this combination of standardized time-domain and frequency-domain features from the five different signal types, a wide range of process information is made available for TCM analysis.

To create the simulated process signals for the noisy training study, the data from Experiments 2B and 3C is replicated four times before features extraction, and different levels of White Gaussian noise signals are additively combined with it to create noisy signals. Five levels of noise are to be assessed, including the unaltered data, and are listed in Table 3.6.

Table 3.6 Additive white Gaussian noise levels

Dataset Label	Description
L0	Original data only (no noise added)
L1	AWGN - SNR25
L2	AWGN - SNR20
L3	AWGN - SNR15
L4	AWGN - SNR10

The levels of noise added are determined by Signal-to-Noise-Ratio (SNR) values, which are selected based on successful noisy training studies conducted in the speech recognition field using sound [33, 157]. After the new signals are created, the 79 ML features can then be extracted for ML analysis.

3.3 Feature Selection

Although access to a large number of features can make more information available to a ML algorithm, it can also distract it from the most important features and hurt the prediction performance [13, 32, 50]. This effect depends on the ML model being used, as well as the specific features' levels of redundancy, dependence on each other, and correlation to the target variable [13, 32]. In addition, calculating and processing a higher number of features increases a TCM system's computation time, and as a result, reduces its feasibility for in-process monitoring.

To determine the optimal number and subset of variables to be used for the analysis, four feature selection techniques are evaluated, including two supervised and two unsupervised feature selection methods. A supervised feature selection method, such as the Recursive Feature Elimination technique used with both SVM and a RF models for this study, takes a ML algorithm's performance changes into account as it searches for the optimal feature subset. Unsupervised

feature selection methods, such as the Mutual Information and correlation techniques assessed here, select features without knowledge of a ML model's performance.

The Recursive Feature Elimination (RFE) technique selects an optimal feature subset size, as well as the features to be used, by initially running a ML model using all features available, and recursively eliminating the next least-important feature and repeating the process [158]. RFE using the ML model's cross validation score for performance evaluation during the process is known as RFECV. Using RFECV, the combination of features which achieves the highest CV score is given as the final feature selection result [159]. RFE can only be run using a ML model which calculates feature rankings, weights, or another measure of feature importance. RFECV using a SVM model (SVM-RFECV) can be done using the algorithm's calculated feature weights, while RFECV using a random forest model (RF-RFECV) can be done using importances based on node impurities and the probability of reaching each node.

The Mutual Information (MI) method [160] calculates the mutual information coefficient, also known as the information gain, for each feature compared to the target variable (in this case, the tool wear level). This MI coefficient represents the uncertainty reduction for one feature when the other feature's value is known, and can take into account several types of potential relationships between features such as patterns between their means, variances, or higher moments [161]. The features with the highest MI coefficients are selected for use, and the number of features to be selected must be input by the user.

Finally, the Correlation-based Feature Selection method (CFS) [162] is also applied, in which a coefficient is calculated based on both the correlation between a feature and the target variable, as well as the intercorrelations between the feature and the other features under consideration. A high correlation to the target variable and low correlations to the other features

will result in the feature being selected for the final feature subset [13]. Generally, the cutoff value for the correlation coefficient is input by the user.

As SVM-RFECV [50, 158, 159], RF-RFECV [163], MI [158, 164], and CFS [13, 165] have all been shown to be promising for TCM feature selection, all four methods will be conducted using the nine ML methods investigated. Through this, it will be assessed how the various ML feature selection method options affect TCM system performance, including that of the ensemble ML systems, for which the feature selection methods have not been compared previously.

3.4 ML Model Evaluation

In order to answer the research questions, the nine selected ML models (DT, SVM, kNN, ANN, EHV, ESV, Stacked SVM, RF, and ET) are evaluated by multiple performance metrics, and are run using several training – testing combinations.

3.4.1 ML Performance Metrics

In order to evaluate the ML models' performance effectively, several metrics are measured.

3.4.1.1 Direct ML Performance Metrics

For the transferability studies, the models' training and testing is repeated 20 times and the classifications' mean accuracy (equal to the weighted recall), accuracy standard deviation, macro-averaged recall, weighted precision, macro-averaged precision, weighted F1 score, and macro-averaged F1 score are recorded. The formulas for the accuracy, recall, precision, and F1 score for each class are shown in Equations 3.1 – 3.4, and depend on that class' true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP} + \boldsymbol{FN}}{\boldsymbol{TP} + \boldsymbol{FP} + \boldsymbol{TN} + \boldsymbol{FN}} \quad (3.1)$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \quad (3.2)$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \quad (3.3)$$

$$\boldsymbol{F1\ Score} = \boldsymbol{2} * \frac{\boldsymbol{Precision} * \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}} \quad (3.4)$$

Each of these metrics measures a different aspect of a model's performance for each class: accuracy measures the proportion of all samples the algorithm classifies correctly; recall measures the proportion of actual positives which are classified correctly; precision measures the proportion of predicted positives which are classified correctly; and the F1 score combines the recall and precision into a single metric.

For multi-class ML classification problems, there are multiple ways of averaging the performance metrics for the different classes in order to get a single value. Weighted averaging gives each ML sample an equal impact on the final performance metric, by averaging the metric for each class using a weight based on the number of samples in that class. Macro averaging, on the other hand, directly averages the classes' metric values without any weighting. If the class sizes are imbalanced, this results in the impact of the smaller classes being raised to equal that of the larger classes. In this study, both averaging results for the recall, precision, and F1 score are evaluated.

3.4.1.2 Cross Validation Metrics

Two other performance metrics are also examined in this study, in order to get a quick overview of a model's performance across a group of data. K-Fold Cross Validation is a commonly applied metric for TCM ML analysis and hyperparameter tuning, and it has shown several advantages [13, 25, 54, 106, 107, 166-168]. As it requires several ML runs to be conducted using different subsets of an overall dataset, it can achieve more accurate performance predictions than through the use of only one training and one testing set. The final K-Fold CV score, the average score for these k ML runs, generally achieves a lower variance than other direct metrics, especially when the number of folds is increased. In addition, the changing training and testing sets helps K-Fold CV detect overfitting in a developed model, as well as evaluate overfit models more accurately than with other metrics. This measure provides a good estimate of how well a ML model would perform when trained and tested on random non-overlapping subsets of a dataset.

However, this approach has a significant disadvantage for some TCM applications: it considers each ML sample to be completely independent from other samples, and therefore it does not take relationships between the samples into account during the k folds' selection. For TCM studies in which multiple distinct machining or environmental conditions are used for experiments and the data is combined for CV, the use of K-Fold CV results in training and testing being done using different samples from the same experiment. This would not be done in realistic tool condition classification applications, where a model would be pre-trained before an experiment starts [16, 169]. Due to the significant differences in tool wear rates and patterns between experiments even if the same experimental parameters are used [19, 20], the inclusion of samples from one experiment in both the training and testing datasets is likely to provide artificially high model performance measures [32, 170].

The Leave-One-Group-Out Cross Validation (LOGO-CV) technique, also known as Leave-One-Cluster-Out CV, is designed to address this issue: instead of selecting the groups to be used for model training and testing based on the individual samples, this algorithm allows samples to be labeled and grouped together according to relationships between them [171- 173]. As a result, for this application each tool life experiment can be considered its own group, preventing samples from one experiment from being present in both the training and testing groups at any point in the CV process. Used in this manner, LOGO-CV provides an averaged measure of how well a model performs when it is trained on data from all available experiments except for one, and then tested on the remaining experiment. While this technique has been shown to be a better measure of model performance than other CV methods in fields such as material discovery [170, 172, 173], it has not previously been applied to TCM. In this study it will be employed in order to gain a better understanding of model performance across different TCM experiments, through the use of more realistic training-testing sets for these situations. However, as k-fold CV has the advantage of being able to split the overall dataset into completely different variations each time it is run, while due to a limited number of groups LOGO-CV is limited in this respect, 10-fold CV will also be used.

3.4.1.3 Model Performance Comparison Metrics

Finally, once the models have been evaluated individually, their performance in relation to each other will be compared. Several metrics are reported for this, including the models' accuracy score standard deviation, the 95% confidence interval of the accuracy mean, bar graphs showing the full range of accuracy scores, and statistical t-test results.

For the measurement of statistical significance of any model performance differences, student's t-tests will be employed using the scores from 20x-repeated 10-fold CV. This is in

agreement with Bouckaert and Frank's study [174] on the optimal methods for comparing ML model performance, which recommended 10x-repeated 10-fold CV due to its low Type 1 error, low Type 2 error, and high replicability scores. The significance of any 20x-repeated LOGO-CV score distribution differences will also be assessed using t-tests.

3.4.2 ML Training, Validation, and Testing Sets

A total of 4,916 original ML samples from the 8 experiments were collected to be used for the TCM analysis. After feature extraction, each ML sample included 79 feature values from its original 1-second-long sound, spindle power, and axial load signals. After the noise-based data augmentation of Experiments 2B and 3C, 4,580 noisy samples were also made available for analysis. For each experiment, ½ of its samples are set aside for the validation dataset, and the other ½ are used for model evaluation. The models' hyperparameter tuning is completed using 5x-repeated 10-fold CV on the entire validation dataset, including data from all 8 experiments. The evaluation dataset is then split into training and testing groups based on the relevant experimental data for each analysis.

CHAPTER 4: RESULTS

The results from the collected process signals, the extracted features, the feature selection techniques, and the developed machine learning models are detailed in this chapter. A summary of the experimental results is given by Table 4.1.

Table 4.1: Overview of experimental results

Experiment Label	# Passes to 0.3mm Flank Wear	Total Material Removed (mm^3)	Total Machining Time (min)	Total # ML Samples Used	Samples in Wear Level 1	Samples in Wear Level 2	Samples in Wear Level 3
1A	54	19910	9.5	468	36	126	306
2B	49	18067	12.3	614	64	269	281
3C	59	21754	10.4	526	54	304	168
4D	63	23229	15.8	801	78	406	317
5D	73	26916	18.3	924	76	439	409
6C	47	17329	8.3	422	34	229	159
7B	48	17698	12.0	604	99	256	249
8A	59	21754	10.4	527	40	250	237
Total	452	166657	96.9	4886	481	2279	2126

It is noted that for each of the experiments, significantly fewer samples were collected for wear level 1 than for wear levels 2 and 3. As shown in Figure 4.7 and discussed in Section 2.1, this is due to the higher wear rates which are expected in the initial stage of tool wear as the cutting edge dulls to form a more sustainable shape. As this imbalance in class sizes is common in TCM due to wear rate variations, it is important that TCM models are effective for unbalanced datasets.

4.1 Process Signals

The collected sound, spindle power, and axial load signals from the eight tool wear experiments are first pre-processed as described in Section 3.2. For Experiment 8A, the sound signal's raw data from the combined machining pass regions is shown in Figure 4.1.

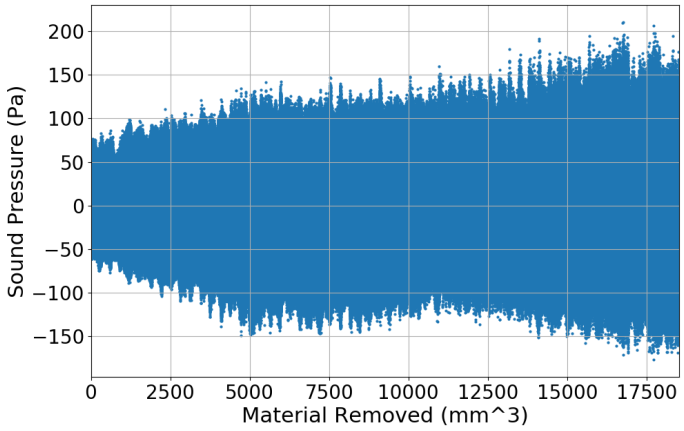
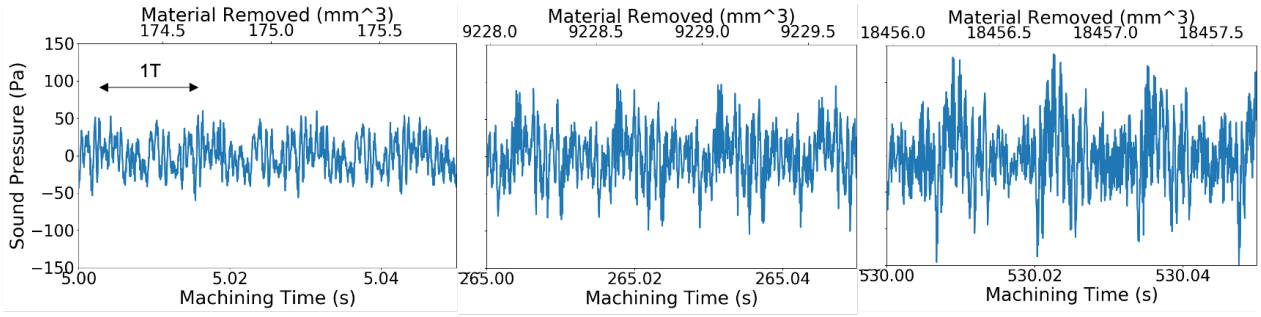


Figure 4.1: Sound pressure data from Experiment 8A

These signals can be viewed in more detail in Figure 4.2, where 0.5-second sound signals are shown at machining times within each of the three tool wear levels. At 5 seconds of machining time (t = 5 sec), VB_{max} = 0.047 mm; at t = 265, VB_{max} = 0.180 mm; and at t = 530, VB_{max} = 0.293 mm.



(a) **(b)** **(c)**

Figure 4.2: Sound signals from Experiment 8A at wear levels (a) 1, (b) 2, and (c) 3, with 1 spindle rotation period (T) displayed for scale

Noticeable increases in the sound signal amplitude can be observed as the tool wear progresses between the three levels, especially at frequencies related to the spindle frequency.

Similarly, the spindle power (SP), x-axis load (XL), y-axis load (YL), and z-axis load (ZL) data, all collected from the machine controller and sampled at 166 Hz, are shown in Figure 4.3 and Figure 4.4. For all five signals collected, peaks at each spindle rotation can be identified, and an increase in amplitude can be observed as the tool becomes more worn.

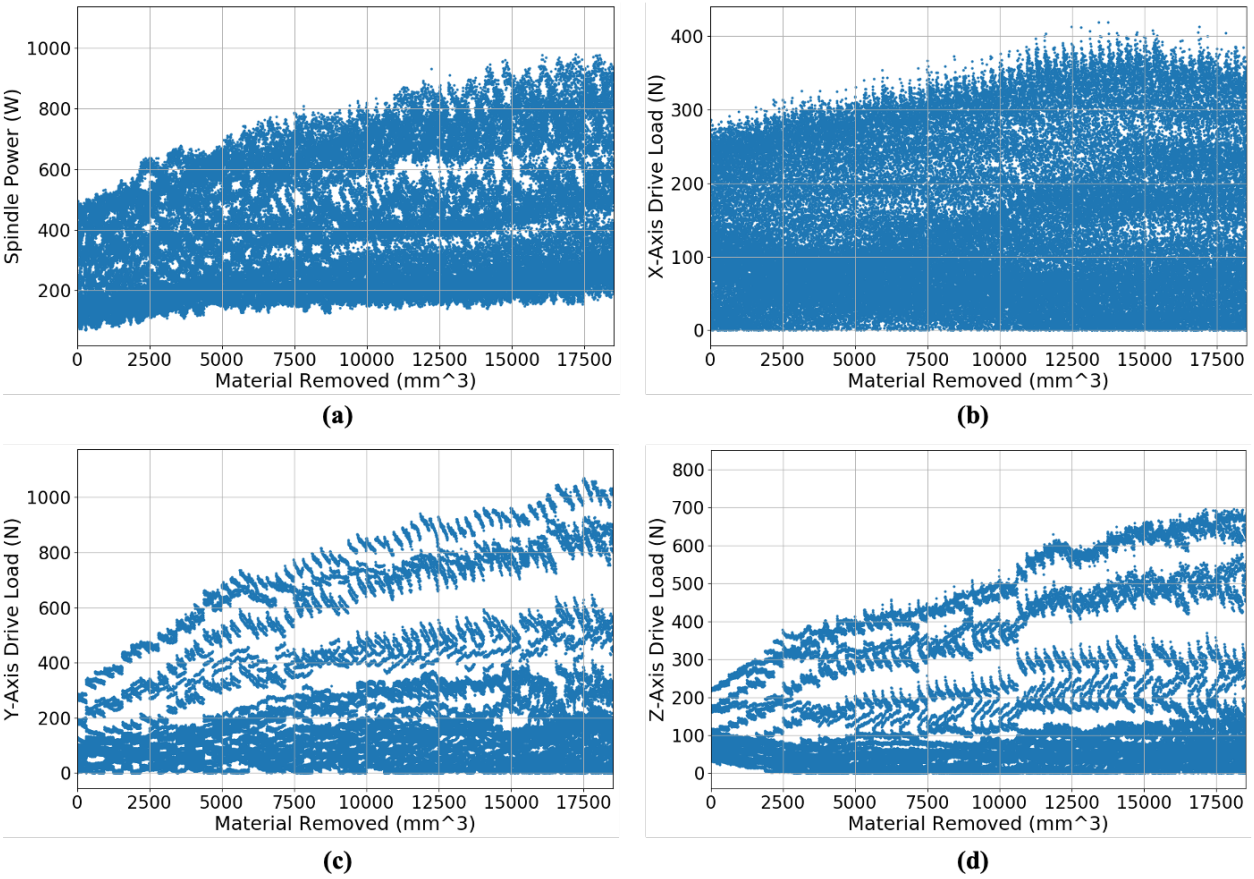


Figure 4.3: Combined machining pass (a) spindle power, (b) x-axis drive load, (c) y-axis drive load, and (d) z-axis drive load data from Experiment 8A

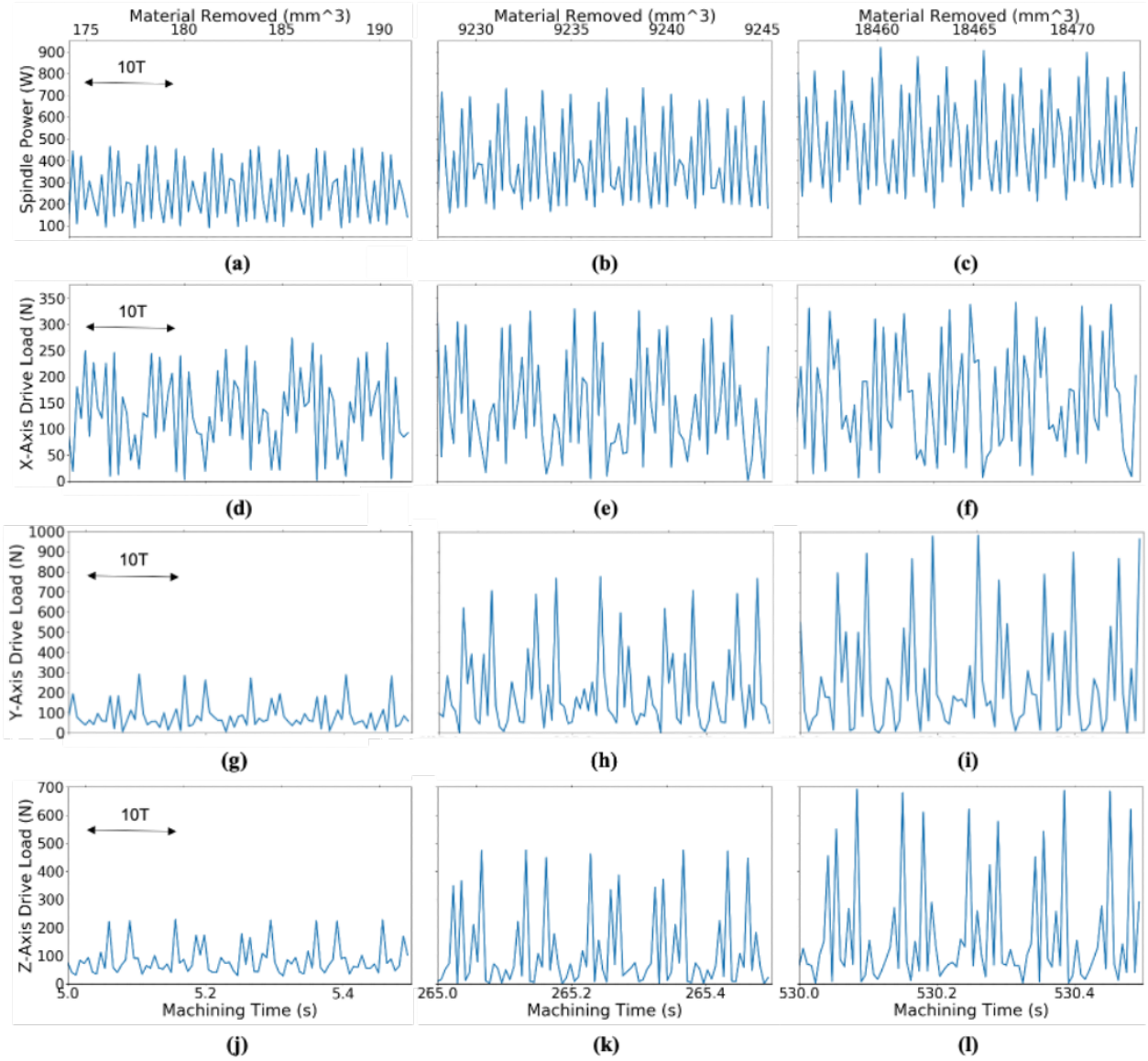


Figure 4.4: Experiment 8A SP signals at wear levels (a) 1, (b) 2, and (c) 3; XL signals at wear levels (d) 1, € 2, and (f) 3; YL signals at wear levels (g) 1, (h) 2, and (i) 3; and ZL signals at wear levels (j) 1, (k) 2, and (l) 3

Figure 4.5 shows the sound signal when converted to the frequency domain using a Discrete Fourier Transform. The same three machining time segments are shown as above, from within each of the three wear levels. There are clear peaks at the spindle frequency and its harmonics, with an increase in those peaks' amplitudes as the tool's wear progresses.

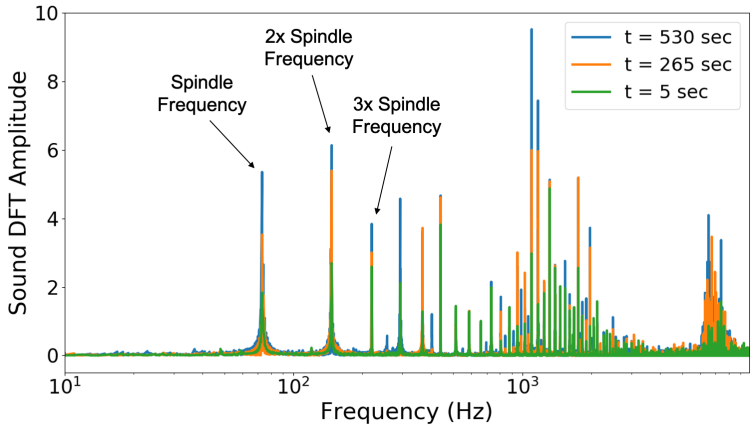
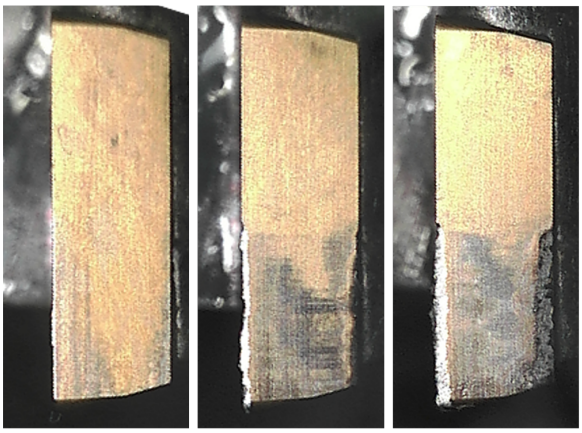


Figure 4.5: DFT of Experiment 8A sound data from within tool wear levels 1 (t = 5 sec), 2 (t = 265 sec), and 3 (t = 530 sec)

Example wear images from the three wear levels are shown in Figure 4.6. Their material removed (MR) values are also listed for reference. The wear progression was observed to be fairly gradual and consistent across the cutting edge for most of each experiment. Near the end of each experiment, the lower cutting edge region began to wear more rapidly than the upper cutting edge, which is consistent with expectations according to ISO8688-2 [55]. No significant chips were identified in any of the 8 experiments at or before the 0.3 mm flank wear cutoff.



(a) **(b)** **(c)**

Figure 4.6: Tool condition deterioration during Experiment 6C. The insert is shown at (a) MR = 940 mm^3 and VB_{max} = 0.075 mm, (b) MR = 8,460 mm^3 and VB_{max} = 0.189 mm, and (c) MR = 14,100 mm^3 and VB_{max} = 0.284 mm

Figure 4.7 shows the flank wear paths followed by each of the 8 tools used. While all of the wear paths appeared to show clear initial wear stages from about 0-0.10 mm VB, and less uniform normal wear stages from approximately 0.10 mm to between 0.25 and 0.30 mm VB, not all of the tools entered the severe wear stage before reaching the 0.3 mm cutoff. This is consistent with previous research, in which the beginning of the severe wear stage can vary widely based on machining conditions and other factors [57, 154, 43, 28].

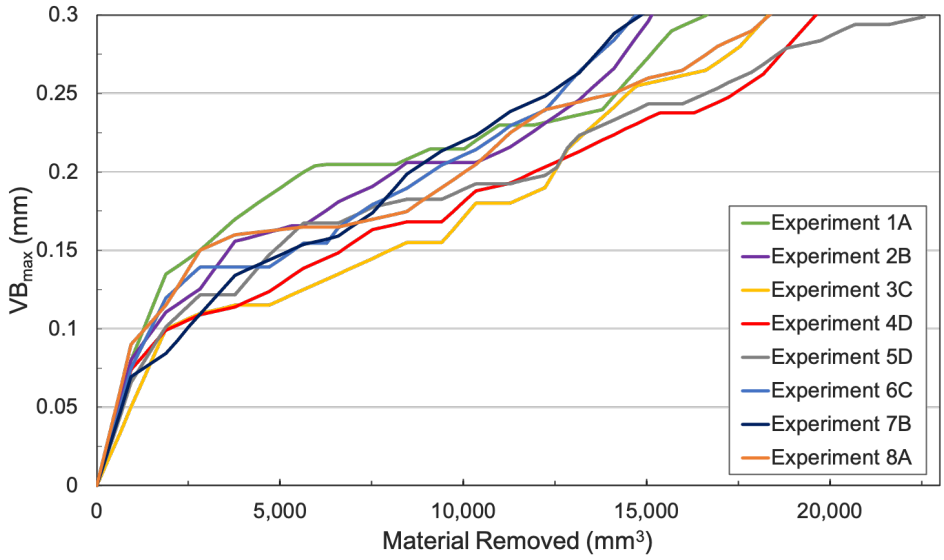


Figure 4.7: Tool flank wear progression for all 8 experiments

Figure 4.7 also shows a few interesting patterns between the cutting parameters used and the tool's overall lifetime from 0–0.3 mm VB. Experiments 4D and 5D, which used the low spindle speed (N) and low feed rate (FR), achieved the longest lifetimes both in terms of machining time and material removed. This is consistent with previous studies which have shown both cutting speed and chip load to be positively correlated to tool wear rates, with cutting speed having the largest effect [2, 5, 82-84]. In addition, the shortest tool lives were observed for parameter set B,

which used a high spindle speed and a low feed rate. This is also consistent with industry knowledge, as this combination of cutting parameters results in the lowest chip load, which can give a high wear rate when the chip load is low enough to result in increased tool rubbing and abrasion on each tool rotation instead of clean cuts into the workpiece. Outside of this rubbing condition, increases in chip load between the other three levels results in increasing wear rates as expected.

4.2 Signal Features

The 79 extracted features show interesting results. Figure 4.8 compares the mean values of all five process signals from Experiment 4D, as well as the tool's wear progression pattern. As with all of the features to be extracted and used for ML, these mean values were calculated from each 1-second segment of collected signal data. In agreement with Experiment 8A, all five process signals showed increases in their mean values as the tool flank wear increased. The sound and spindle power signal mean values even showed similar changes in slope to that of the wear pattern at approximately MR = 2000 mm^3, when the tool transitioned from stage 1 to stage 2 wear.

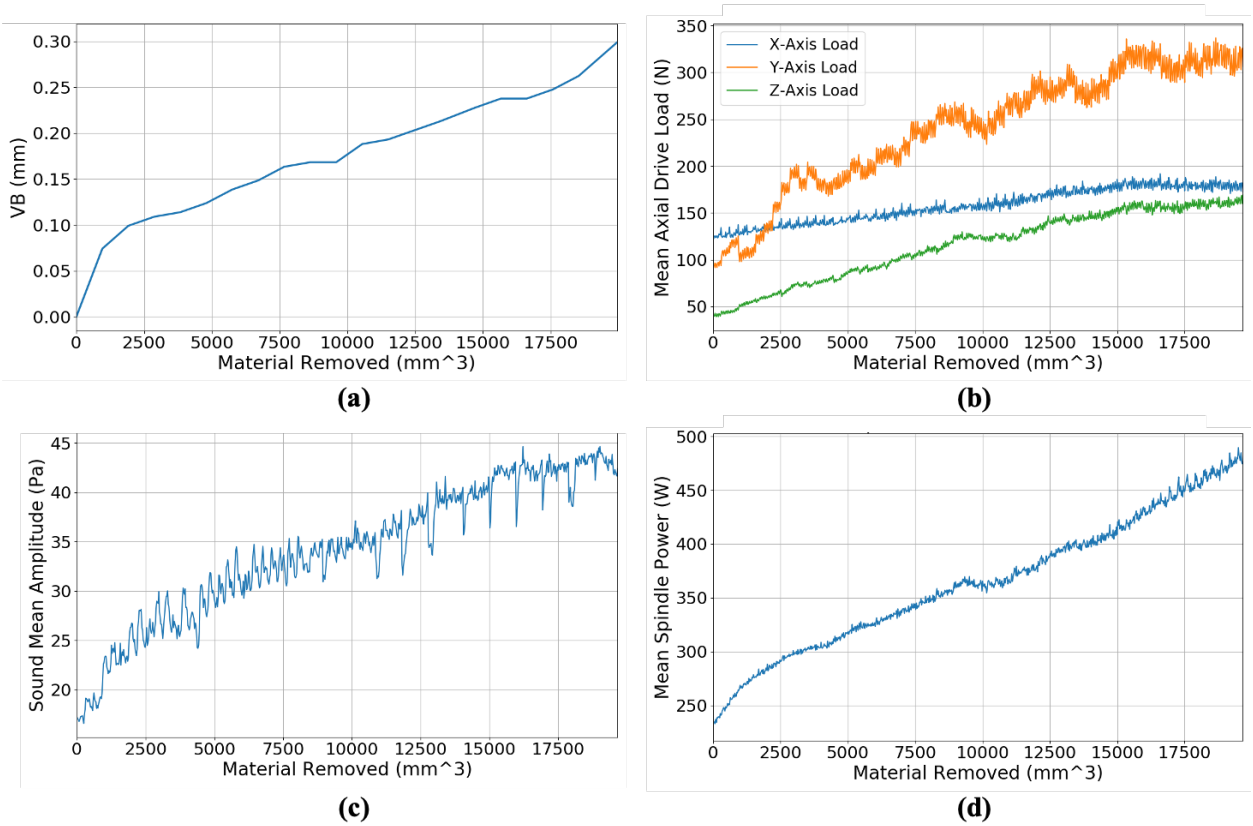


Figure 4.8: Experiment 4D (a) tool flank wear, (b) mean axial load, (c) mean sound pressure amplitude, and (d) mean spindle power experimental results

Several other extracted features also show value increases with the progression of tool wear, although their magnitudes and pattern details may vary between individual experiments. Some of the features which aligned best between the different experiments and were therefore more suitable options for machine learning analysis, are shown in Figure 4.9 and Figure 4.10.

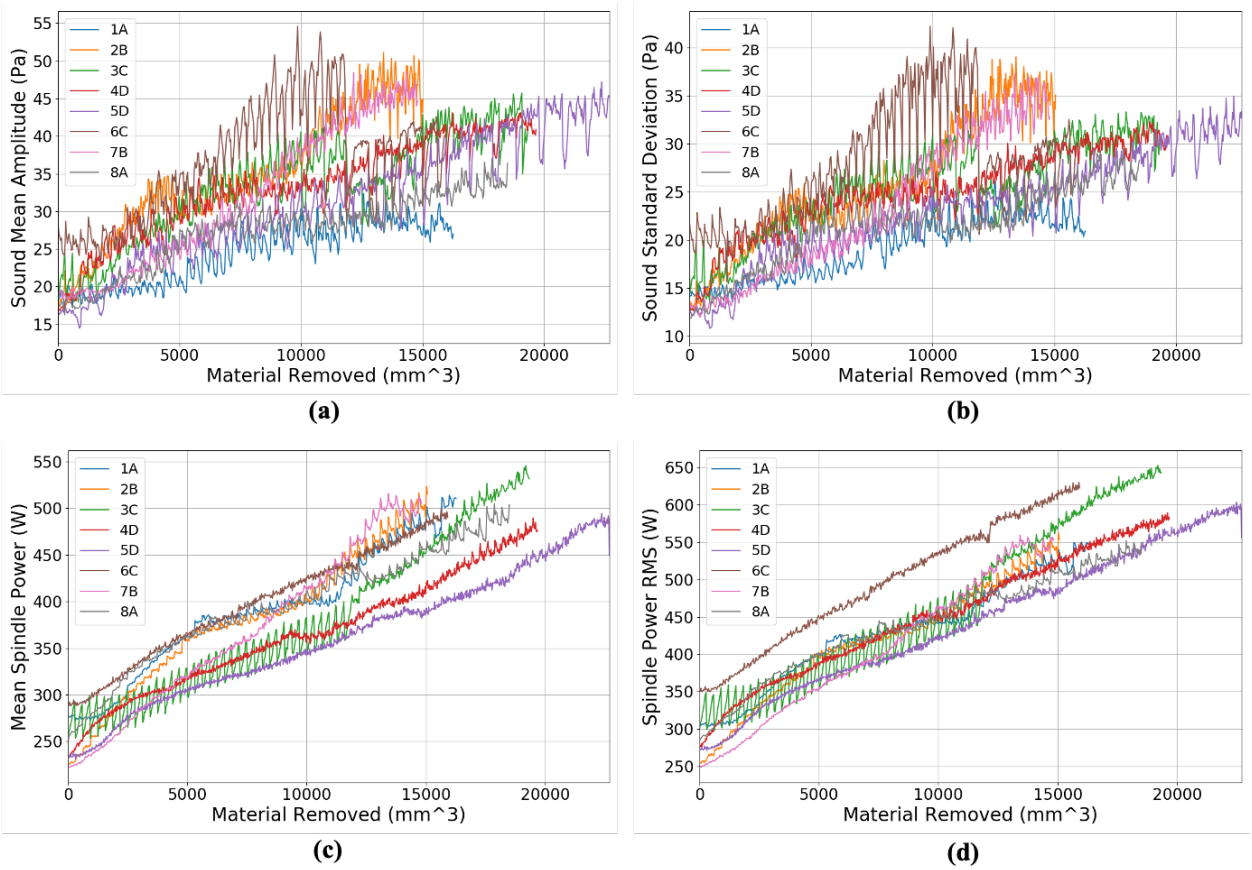


Figure 4.9: All experiments' (a) mean sound signal amplitude, (b) sound signal standard deviation, (c) mean SP, and (d) SP root-mean-square (RMS)

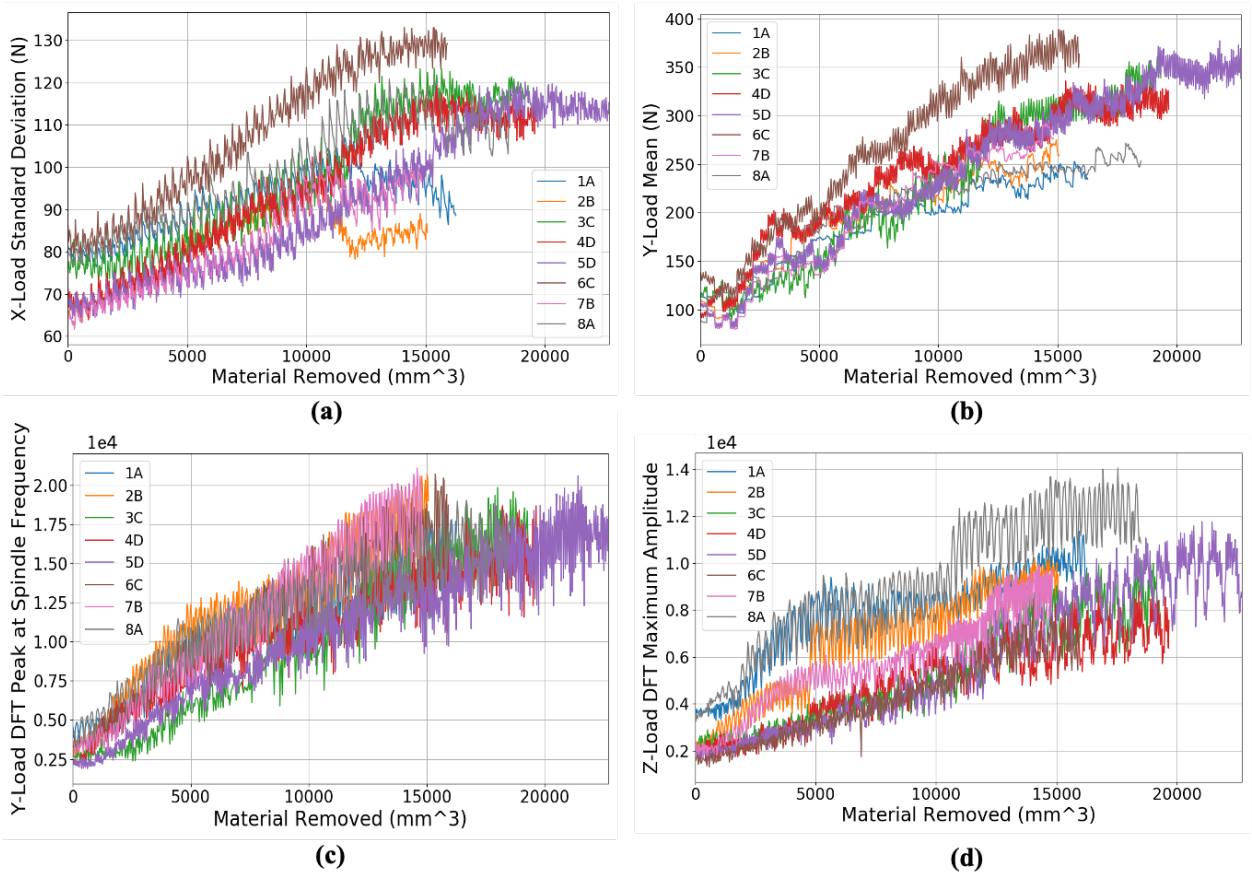
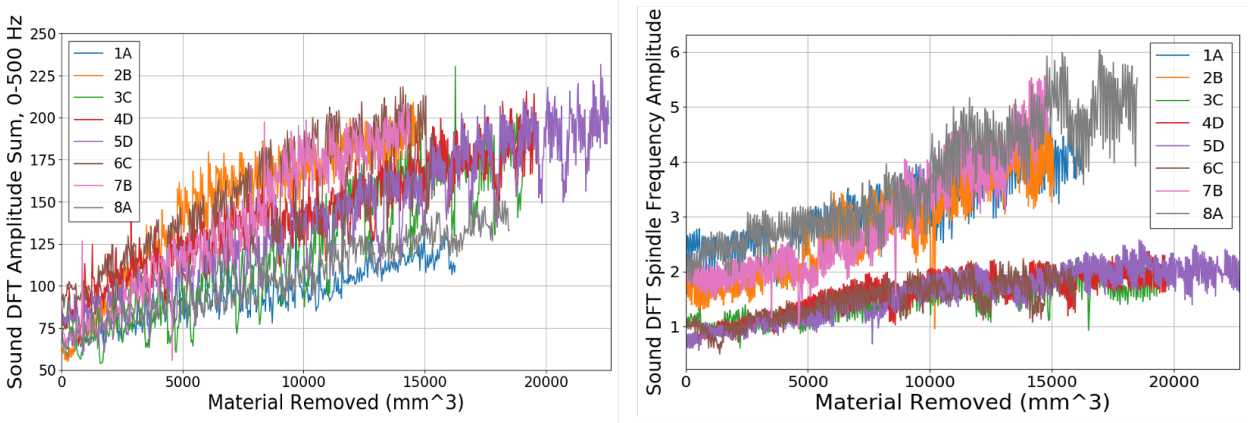


Figure 4.10: All experiments' (a) XL standard deviation, (b) mean YL, (c) YL DFT peak at the spindle frequency, and (d) ZL RMS

In the frequency domain, several frequency bands and individual peaks were evaluated for correlation to tool wear as well as consistency across experiments. For the signals sampled by the machine controller, including the SP and axial loads, the sampling rate of 166 Hz allowed the spindle frequency to be monitored but was not high enough to evaluate its harmonics or any higher frequency bands. However, the DFT amplitude at the spindle frequency showed strong correlation to tool wear in the SP, YL, and ZL signals. For example, Figure 4.10 € shows the patterns followed by the spindle frequency peak amplitude in the y-axis load signals during different experiments, which all increase with wear and overlap significantly.

As the sound signal was sampled at a much higher rate, its frequency bands can be assessed in more depth. From the sound signals, two frequency peaks and four frequency ranges were monitored: the spindle frequency and its second harmonic, the full DFT range (0-9 kHz), a 0-500 Hz low frequency band which encompasses several spindle rotation harmonics, a 4-6 kHz band based on L. C. Lee's characteristic frequency range [94], and a 2.75-3.75 kHz frequency band based on Sadat and Raman's work [93]. The clearest of these were the 0-500 Hz frequency range amplitude sum and the spindle frequency peak amplitude, both shown in Figure 4.11. However, the spindle frequency peak amplitude appears to be highly dependent on the spindle speed (N) used, as the four experiments with N=4400 RPM (parameter sets A & B) follow a separate path from the four which used N=3700 RPM (parameter sets C & D).



(a) **(b)**

Figure 4.11: All experiments' sound DFT (a) 0-500 Hz amplitude sum, and (b) amplitude at the spindle frequency

Although peaks in the 4-6 kHz range were observed in the sound signals, and both the 4-6 kHz and 2.75-3.75 kHz frequency ranges saw increases in amplitude sum for some individual tools' lives, no clear pattern across all cutting parameters used was found for either band. This is not surprising due to the differences between the cutting processes used. On the other hand, the 0-9

kHz amplitude sum and the 2nd harmonic of the spindle frequency showed similar increases and correlations to tool wear as those in Figure 4.11, although less clearly defined. These strong correlations to the spindle frequency harmonics at lower frequency values are more expected for milling processes [29]. All of these calculated features were made available to the feature selection algorithms.

4.3 Feature Selection Results

The four feature selection methods, SVM-RFECV, RF-RFECV, MI, and CFS, run on the validation dataset which includes data from all eight experiments, resulted in the selection of 14, 3, 36, and 20 features respectively. The features selected by each method are listed in Table A.1, as well as summarized by signal type in Table 4.2 below. The features extracted from the z-axis load signals were selected the most frequently overall by the feature selection methods, and the x-axis load signal's features were selected the least.

Table 4.2: Overview of feature selection results

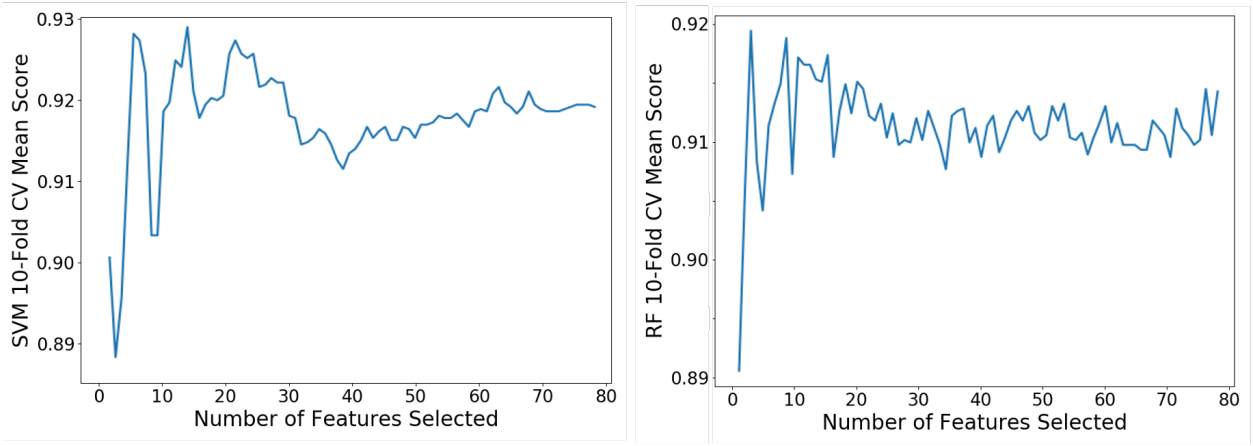
	Features Selected by various Selection Methods					% Features Selected from each Signal, across all 4 Feature Selection Methods
Feature Selection Method Used:	**No Feature Selection**	**SVM-RFECV**	**RF-RFECV**	**MI**	**CFS**	
Total Number of Features Selected:	79	14	3	36	20	
# Features Selected from Sound Signal	19	5	0	5	0	13.70%
# Features Selected from Spindle Power Signal	15	2	1	9	4	21.92%
# Features Selected from X-Load Signal	15	2	0	4	1	9.59%
# Features Selected from Y-Load Signal	15	3	0	9	7	26.03%
# Features Selected from Z-Load Signal	15	2	2	9	8	28.77%

The most frequently selected individual features are also listed in Table 4.3 for reference. This list gives an idea of the features which are the most highly correlated to tool wear, as well as the least dependent on, or correlated to, the other signal features.

Table 4.3: Most frequently selected features

Features:	Occurrances:
SP TD mean	4
SP TD RMS	3
XL TD standard deviation	3
YL TD mean	3
YL TD maximum	3
ZL TD mean	3
ZL TD standard deviation	3
ZL TD RMS	3
Sound TD mean	2
Sound TD standard deviation	2
Sound FD amplitude sum 0-9kHz	2
SP FD maximum amplitude	2
SP FD peak at spindle frequency	2
YL TD standard deviation	2
YL TD RMS	2
YL TD range	2
YL FD maximum amplitude	2
YL FD peak at spindle frequency	2
ZL TD median	2
ZL TD maximum	2
ZL TD range	2
ZL FD maximum amplitude	2
ZL FD peak at spindle frequency	2
ZL FD amplitude sum 0-83Hz	2

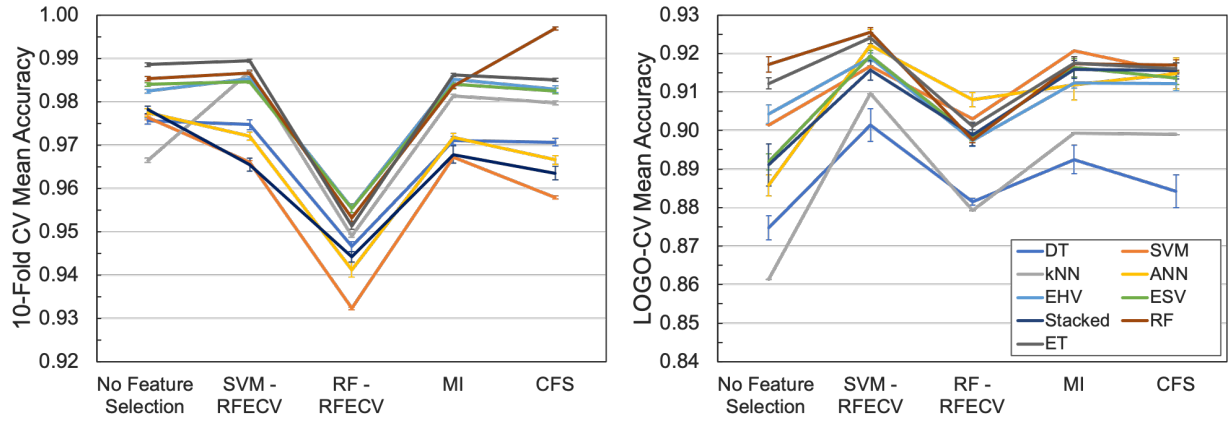
Figure 4.12 shows the recursive feature elimination process followed by the SVM- and RF-based feature selection methods. As described in Section 3.3, the process begins on the right end of the plots, with all 79 features used for cross validation scoring, and moves toward the left as one feature at a time is eliminated.



(a) **(b)**

Figure 4.12: RFECV feature selection process, using (a) SVM, and (b) RF models for evaluation

The effect of the different feature selection (FS) methods on the performance of the nine machine learning algorithms studied in this work are shown in Figure 4.13. Both 10-fold cross validation and LOGO-CV performance metrics are used in order to get an idea of how well each set of features performs on the dataset as a whole, as well as how the feature selection methods impact model transferability to data from new unseen experiments.



(a) **(b)**

Figure 4.13: Feature selection methods' impact on model performance. Average accuracy scores for both (a) 20x-repeated 10-fold CV, and (b) 20x-repeated LOGO-CV are shown, with error bars depicting the 95% confidence interval ranges

The SVM-RFECV method results in the highest average LOGO-CV score across the nine models, at 91.7%, as well as a 10-fold CV average score of 97.9% which is tied as the top score with No FS. Table 4.4 shows the statistical analysis results for the feature selection comparison.

Table 4.4: Statistical analysis of FS method scores

Using 20x-Repeated 10-fold CV Scores	FS Method 1: Chosen Method	SVM-RFECV				
	FS Method 2: Comparison Method	No FS	SVM-	RF-RFECV	MI	CFS
	t-value:	-1.19		158.27	3.21	18.18
	p-value:	2.53E-01		3.56E-24	6.25E-03	3.90E-11
	p-value < 0.05 ?	No		Yes	Yes	Yes
Using 20x-Repeated LOGO-CV Scores	FS Method 1: Chosen Method	SVM-RFECV				
	FS Method 2: Comparison Method	No FS	RFECV	RF-RFECV	MI	CFS
	t-value:	33.74		40.00	8.15	10.32
	p-value:	8.22E-15		7.76E-16	1.10E-06	6.33E-08
	p-value < 0.05 ?	Yes		Yes	Yes	Yes

While there is not a significant difference between the 10-fold CV scores when the SVM-RFECV and the No FS feature selection methods are used, the SVM-RFECV technique resulted in significantly higher scores than all FS methods in every other measure. This is consistent with Li et al.'s study [158], in which SVM-RFE was found to result in higher TCM accuracy scores than both MI and a Fisher Score technique when combined with a least-squares SVM classification model. For these reasons, SVM-RFECV is selected for feature selection and its subset of 14 features are used for the remainder of the study.

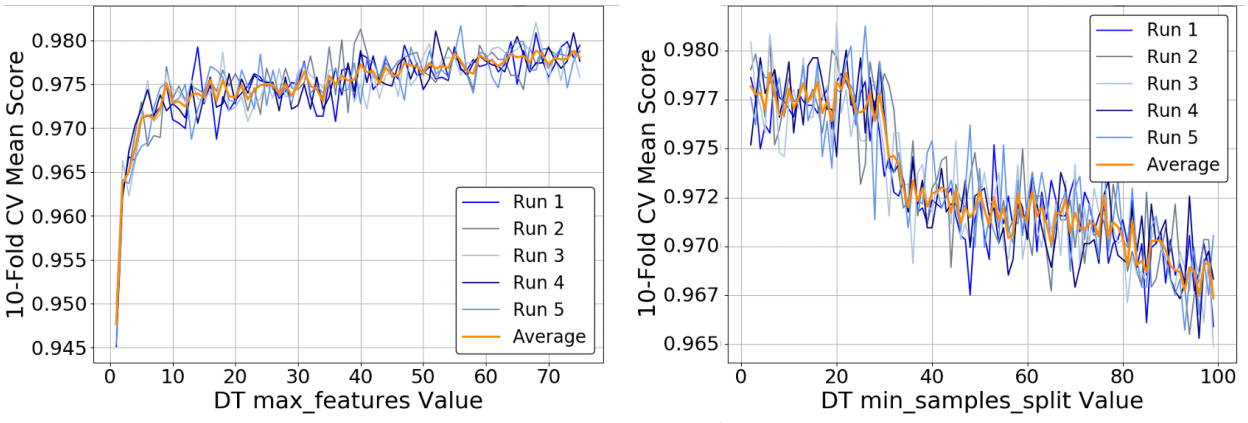
4.4 Developed ML Models

Using the feature subset selected through SVM-RFECV, the nine ML models are then optimized by tuning their hyperparameters to fit this application. This process and the models'

initial results are presented in Sections 4.4.1 and 4.4.2. The developed models' computation times are also assessed in Section 4.4.3.

4.4.1 Model Hyperparameter Optimization

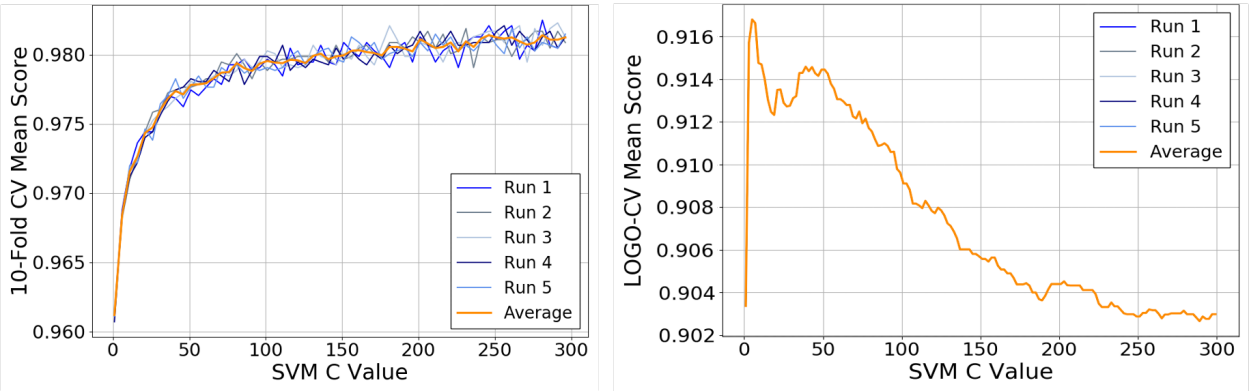
The nine ML models' hyperparameters are tuned to adjust the algorithms to the TCM application, using a combination of parameter grid searches, 5x repeated 10-fold CV, and 5x repeated LOGO-CV all run on the entire validation data subset. 10-fold CV is generally a good performance metric for hyperparameter tuning since its several unique training and testing groupings, especially when run and averaged five times, help prevent models from becoming overfit on any specific training sets. For the decision tree model, Figure 4.14 shows the tuning process for two hyperparameters: the maximum number of features considered at a time, and the minimum number of samples needed for an internal node to split. Values of "None" and 2 were selected for these parameters, no maximum tree depth was set, and a Gini impurity criterion was selected for measuring the node split quality.



(a) **(b)**

Figure 4.14: Tuning of DT (a) max. features and (b) min. samples to split hyperparameters

As direct SVM classification algorithms are designed for binary classification, a one-vs-one scheme is used for this application in order to classify between the multiple wear levels. A radial basis function kernel is selected for the SVM using a grid search, and Figure 4.15 shows the process of optimizing the regularization parameter C. As this parameter controls the model's penalty for misclassifications in the training data, it must be selected correctly in order to balance the model's prediction accuracy for data very similar to the training data, with the model's generalizability to new and different data. For these reasons, both 10-fold CV and LOGO-CV are used to tune this parameter, and an optimal value of C = 5 is selected.



(a) **(b)**

Figure 4.15: Tuning of SVM C regularization parameter, using (a) 10-fold CV, and (b) LOGO-CV

Figure 4.16 shows the vector regions calculated by the tuned SVM classifier when it is trained on the evaluation data for Experiments 1-4. The data points shown are this Experiment 1-4 data, colored by the samples' true wear classes to show how closely the regions fit the training dataset. For visualization, only two features' data were assessed at a time. It is observed that while some of the features have fairly clear cutoffs between the wear levels, such as the mean SP and the ZL standard deviation (SD), others, such as the XL SD and the sound signal mean, are less

clear. Overall, however, the model appears to fit the regions to the data well and without overfitting.

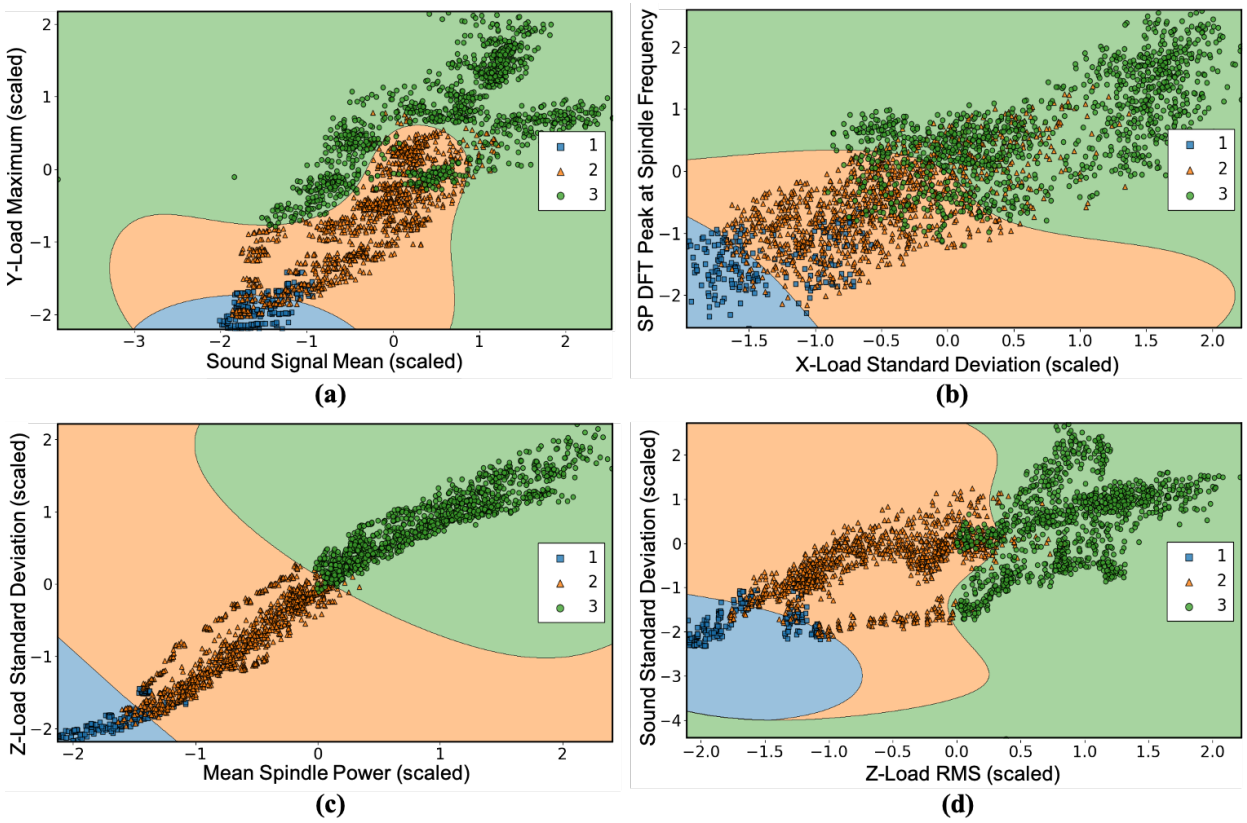
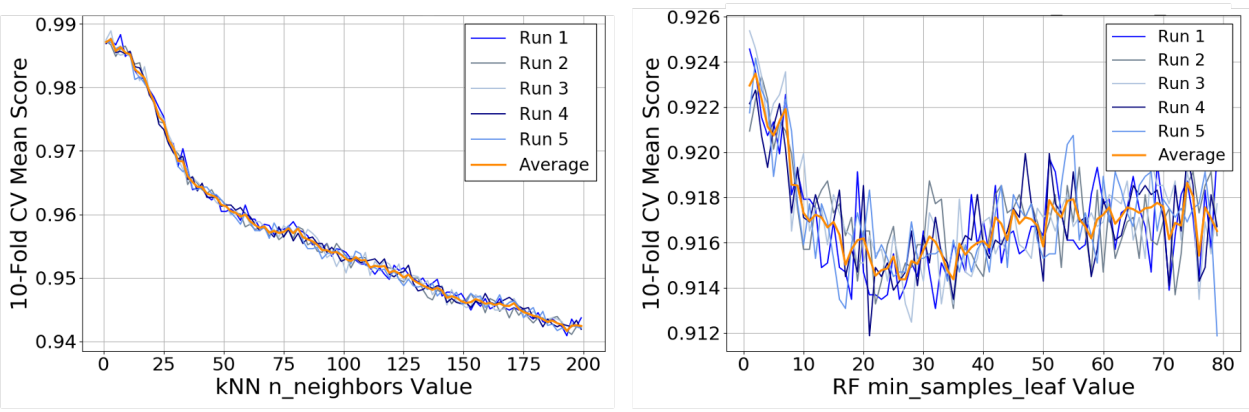


Figure 4.16: SVM vector region plots, using only (a) the mean sound signal and the YL maximum, (b) the XL SD and the SP DFT amplitude at the spindle frequency, (c) the mean SP and the ZL SD, and (d) the ZL RMS and the sound SD

Figure 4.17 shows the optimization process for two kNN and RF hyperparameters, which results in the number of neighbors ("k") for the kNN being set to 3, and the minimum number of samples at each leaf node for the RF being set to 1. Other settings chosen for the kNN included Euclidean distance measures and uniform sample weights. A "forest" of 100 trees is used for the RF, and the other selected parameters match those described for the individual DT model. The extra-trees algorithm also uses 100 trees and this same hyperparameter set.



(a) **(b)**

Figure 4.17: Tuning of (a) the kNN number of neighbors used, and (b) the RF minimum leaf size

The other models are tuned similarly. To tune the multilayer perceptron neural network, several intricate grid searches were utilized to find the optimal activation function, weight optimization solver, L2 regularization parameter, and hidden layer sizes. For these hyperparameters, a rectified linear unit function, a stochastic gradient-based optimizer, a value of 0.0001, and hidden layer sizes of 6, 30, and 37 were found to give the best repeated 10-fold CV results.

For the EHV and ESV ensemble models, the base classifier weights were set according to the base models' 10-fold CV scores using the training data. This allows ties between the four base models' votes to be avoided, and gives slightly more weight to better-performing base models. The DT, SVM, kNN, and ANN algorithms were used as base models. The stacked generalization SVM ensemble method used a data splitting strategy similar to 10-fold CV for determining the training sets for the base models and the SVM meta-learner, as they should not be trained on the same data. While the stacked SVM model was based on the same four base learners as the two voting ensembles, no classifier weights were used for this ensemble method. The SVM meta-

learner's C regularization parameter was set to 1. The use of these nine ML techniques' sets of optimized hyperparameters helps the models better fit the specific TCM application, as well as achieve higher prediction accuracy scores.

4.4.2 ML Initial Results

Using data from all 8 experiments, the overall performances of each ML model are displayed as box plots in Figure 4.18 in terms of their 10-fold CV and LOGO-CV scores. As shown, the homogeneous ensemble models performed the best in both metrics. The statistical significance of the performance differences between these top-ranking models and the other models is assessed in Table 4.5 using t-tests. It is determined that the extra-trees model performs significantly better than all other models in 10-fold CV, and that the random forest performs significantly better in the LOGO-CV analysis than all models except for ET. The decision tree and the neural network show the highest score variability, especially in the LOGO-CV scores. The stacked generalization SVM ensemble model also showed relatively high variability in its 10-fold CV scores.

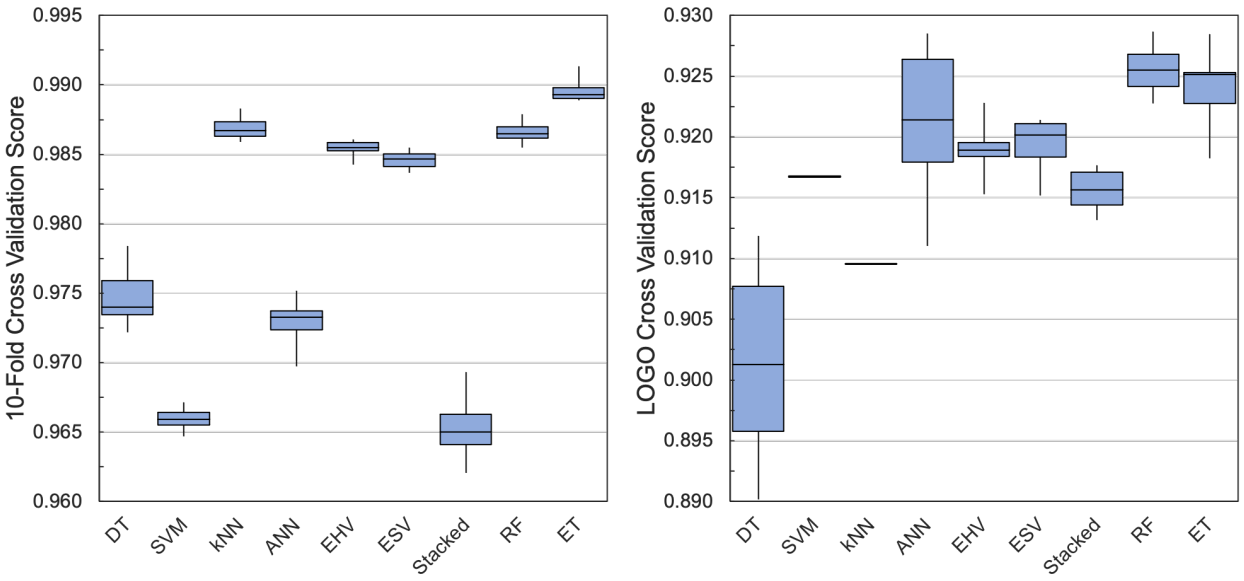


Figure 4.18: Box plots for all ML models, using 20x-repeated (a) 10-fold CV and (b) LOGO-CV

Table 4.5: Statistical analysis of model performance differences

Using 20x-Repeated 10-fold CV Scores	Model 1: 5x2CV Best-Performing Model	Extra-Trees (ET)								
	Model 2: Comparison Model	DT	SVM	kNN	ANN	EHV	ESV	Stacked SVM	RF	ET
	t-value:	34.30	108.61	10.42	36.87	15.77	18.96	38.17	12.08	
	p-value:	1.52E-22	5.82E-35	5.68E-12	5.79E-22	7.96E-14	1.55E-15	6.92E-21	1.28E-12	
	p-value < 0.05 ?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	
Using 20x-Repeated LOGO-CV Scores	Model 1: LOGO Best-Performing Model	Random Forest (RF)								
	Model 2: Comparison Model	DT	SVM	kNN	ANN	EHV	ESV	Stacked SVM	RF	ET
	t-value:	9.03	16.52	39.05	2.55	8.66	7.61	12.72		1.91
	p-value:	2.409E-09	5.798E-15	3.335E-29	1.807E-02	1.061E-08	9.896E-08	2.467E-11		6.653E-02
	p-value < 0.05 ?	Yes	Yes	Yes	Yes	Yes	Yes	Yes		No

The results from the machining conditions transferability study are shown in Table A.2 and discussed in more detail in Section 5.2. Confusion matrices are shown for each of the individual ML models in Figure 4.19, as well as for the ensemble ML models in Figure 4.20, for the runs in which they were trained on data from Experiments 1-4 and tested on data from Experiments 5-8.

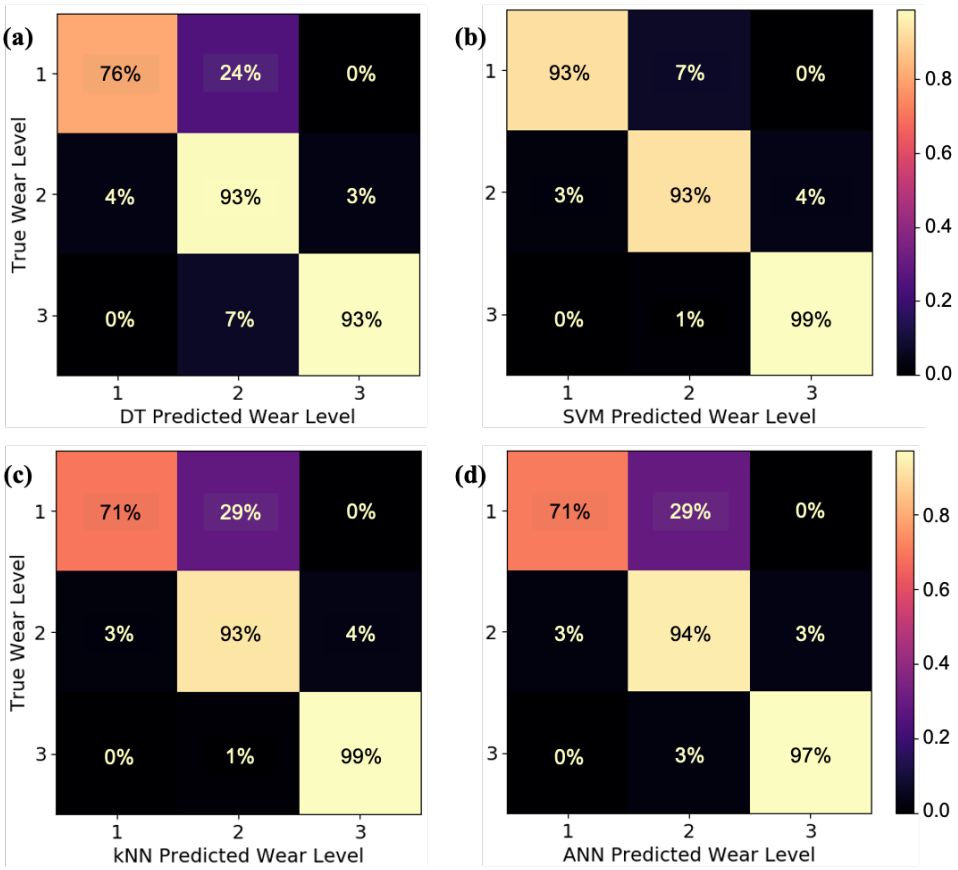


Figure 4.19: Confusion matrices for (a) DT, (b) SVM, (c) kNN, and (d) ANN when trained on data from Experiments 1-4 and tested on data from Experiments 5-8

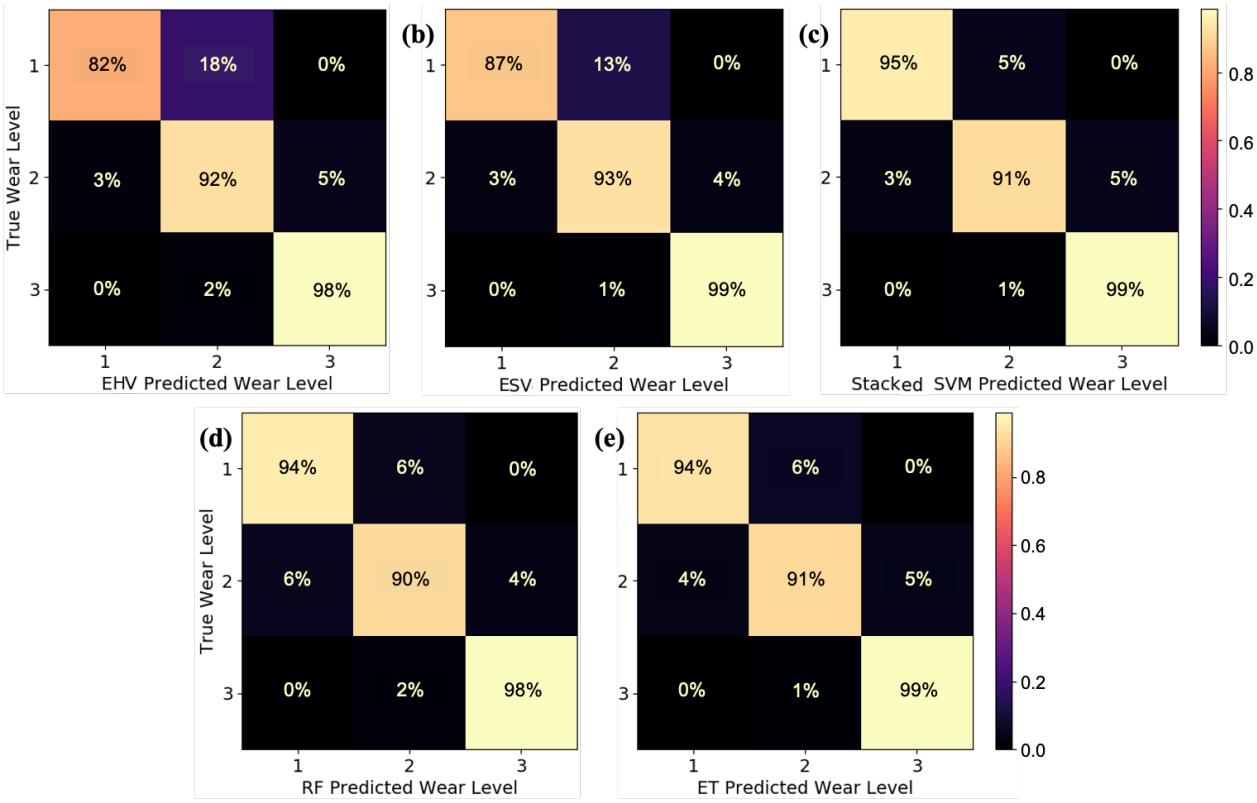


Figure 4.20: Confusion matrices for (a) EHV, (b) ESV, (c) Stacked SVM, (d) RF, and € ET when trained on data from Experiments 1-4 and tested on data from Experiments 5-8

The percentages shown in the confusion matrices are the percentages of the total true samples for each wear level which are predicted to be at the specified level. Notably, the percentages shown along the downward diagonal give the models' prediction accuracies for each of the three wear classes. Overall, the models can be observed to perform better on the second and third wear levels than they do on the first wear level. This is a fairly consistent pattern across the various ML runs and training/testing sets. However, the ensemble methods generally achieve much higher prediction accuracies for the first wear level than the individual models do, showing that the ensemble models have higher robustness to unbalanced datasets. The differences in model performance between the individual models and the ensemble models will be discussed in more detail in Section 5.1.1.2.

4.4.3 Computation Time

The computation times for each model were compared as well, and are shown in Table 4.6. These model classification times ranged from 0.0006 ms per ML sample for the decision tree, up to 0.1184 ms per sample for the hard voting ensemble classifier. While it is expected that the ensemble models will require more classification computation time due to their more complex structures, four out of the five ensemble models achieved lower computation times than the kNN individual model. These results are all based on the 14-feature subset chosen by the SVM-RFECV feature selection method.

Table 4.6: Feature extraction and classification computation times per ML sample

ML Model	Time to Calculate Sound Features (ms)	Time to Calculate Controller Features (ms)	Average ML Classification Time (ms)	Total Computation Time during Process, per Sample (ms)
DT	195.4545	0.9567	0.0006	196.4118
SVM	195.4545	0.9567	0.0211	196.4323
kNN	195.4545	0.9567	0.0732	196.4844
ANN	195.4545	0.9567	0.0025	196.4137
EHV	195.4545	0.9567	0.1184	196.5296
ESV	195.4545	0.9567	0.0488	196.4600
Stacked SVM	195.4545	0.9567	0.0656	196.4768
RF	195.4545	0.9567	0.0316	196.4428
ET	195.4545	0.9567	0.0323	196.4435

Any of these model classification times only make up a small fraction of the amount of computation time necessary for full in-process tool wear monitoring using the proposed system, however. As shown in Table 4.6, the computation time required for the sound signal's feature extraction makes up the majority of the total computation time required, due to the high sampling rate used. However, the calculated total time is still far below the feasibility requirement, which is that the total time to process and classify one ML sample must be less than one second, as each ML sample spans one second of process data. Even if the EHV model is used, the total computation

time for each ML sample comes out to only 0.197 seconds, which is much less than 1 second and is therefore acceptable for in-process TCM. These results demonstrate the suitability of the proposed systems for TCM, as well as show that the increased complexity of ensemble ML techniques is not a hindrance to TCM system feasibility.

CHAPTER 5: DISCUSSION

Despite high industry demand, for a TCM system to be adopted by manufacturers it must meet stringent practicality, accuracy, and generalizability requirements. While the proposed TCM system meets these practicality requirements by utilizing only cost-effective, non-intrusive, and easy-to-install sensors and equipment, the effects of the selected ensemble ML algorithms and configurations on a system's accuracy, and generalizability are to be assessed in this section. In addition, the noisy training method of improving model generalization in situations with limited training data will be investigated.

5.1 ML Optimal Configuration

In order to determine the optimal ML model configurations and settings to be used for this specific TCM application, the nine model types, the effects of heterogeneous and homogeneous ensemble ML configurations, and the impact of different tool wear classification resolutions are evaluated.

5.1.1 Model Performance Comparison

5.1.1.1 Evaluation of All Models

Figure 4.13 and Figure 4.18, discussed earlier, show the 10-fold CV and LOGO-CV scores for each of the nine ML models using data from all eight experiments. In Figure 4.13, it is observed that across the different feature subsets which are investigated, the two homogeneous ensemble models, RF and ET, fairly consistently achieve the highest 10-fold CV and LOGO-CV performance metrics. The two ensemble voting schemes, EHV and ESV, also generally perform

well compared to the remaining models. Figure 4.18 confirms these results for the selected SVM-RFECV-based subset of features, although it also shows the kNN achieving comparable 10-fold CV scores, and the ANN achieving comparable LOGO-CV scores. The stacked SVM ensemble model gives the lowest 10-fold CV scores and mid-range LOGO-CV scores.

Table A.2, Figure 4.19, and Figure 4.20 give detailed results for each model when they are trained and tested on various experimental subsets of data. From the experimental generalizability runs in Table A.2, the results of which are summarized in Table 5.1, it is observed that on average, the stacked SVM ensemble achieved the highest accuracy scores, with the two voting ensembles and the kNN next. However, across all of the other performance metrics, the top scores are always achieved by the ET or the RF, except for the weighted F1 score which is calculated directly from the weighted recall score. Since the Stacked SVM, on average across the experiment generalizability tests, achieved the highest weighted recall score but a relatively low macro-averaged recall score, this suggests that it usually performed sub-optimally on the data from tool wear level 1, as this level contained the smallest number of true samples. As macro-averaging equalizes the weights of the wear classes, this would increase the importance of the first wear level's data in the recall metric and have the observed effect. In addition, for almost every metric in which the ET model did not score the highest, it scored the second-highest. Overall, when all performance metrics are considered, including the 10-fold CV scores, the LOGO-CV scores, the generalizability results, the models' effectiveness for unbalanced datasets, and the computation times, the extra-trees algorithm was determined to be the best model for this application.

Table 5.1: Summarized performance results for all models, with highest scores highlighted

		ML Model								
	Performance Metric:	DT	SVM	kNN	ANN	EHV	ESV	Stacked SVM	RF	ET
Experiment Generalizability Study Averages	Average Accuracy (Weighted Recall) Score:	0.8362	0.8576	0.8792	0.8532	0.8789	0.8798	0.8835	0.8626	0.8731
	Standard Deviation of Accuracy Score:	0.0391	0.0000	0.0000	0.0444	0.0225	0.0154	0.0175	0.0064	0.0055
	95% Confidence Interval of Accuracy Mean:	0.0242	0.0000	0.0000	0.0275	0.0139	0.0096	0.0109	0.0040	0.0034
	Macro-Averaged Recall Score:	0.7694	0.7781	0.7678	0.7844	0.7739	0.7679	0.7740	0.8068	0.7940
	Weighted Precision Score:	0.8699	0.8981	0.8813	0.8871	0.8918	0.8945	0.8995	0.8971	0.9099
	Macro-Averaged Precision Score:	0.8119	0.8748	0.7768	0.8155	0.8268	0.8403	0.8542	0.8571	0.8912
	Weighted F1 Score:	0.8271	0.8297	0.8678	0.8386	0.8636	0.8641	0.8729	0.8528	0.8612
	Macro-Averaged F1 Score:	0.8119	0.8748	0.7768	0.8155	0.8268	0.8403	0.8840	0.8571	0.8912
CV Scores using all Experiments	10-Fold CV Mean Score:	0.9747	0.9660	0.9869	0.9721	0.9854	0.9846	0.9654	0.9866	0.9895
	10-Fold CV Score Standard Deviation:	0.0018	0.0007	0.0008	0.0015	0.0006	0.0006	0.0022	0.0006	0.0006
	LOGO-CV Mean Score:	0.9014	0.9167	0.9095	0.9222	0.9190	0.9194	0.9156	0.9255	0.9240
	LOGO-CV Score Standard Deviation:	0.0068	0.0000	0.0000	0.0065	0.0018	0.0020	0.0015	0.0018	0.0024
Tests within 1 Experiment	Average Accuracy for Same Experiment:	0.9736	0.9689	0.9803	0.9757	0.9812	0.9808	0.9712	0.9833	0.9883
	Same Experiment Accuracy SD:	0.0026	0.0000	0.0000	0.0075	0.0041	0.0033	0.0051	0.0021	0.0021

Across all nine models, the 10-fold CV scores are observed to fall close to the high accuracy scores for models which are trained and tested on separate data from the same experiment, and relatively far from the lower LOGO-CV and experiment generalizability runs in which models were tested on experiments which they had not seen before. This is likely due to the high variability in tool wear rates and patterns which exist under even the most controlled conditions [19, 20, 30]. As a TCM classification model to be used in industry applications would be trained prior to any experiments in which it would be applied, these results confirm the need for the LOGO-CV performance metric or similar model generalizability tests to be used in order for a TCM model's performance to be evaluated realistically.

5.1.1.2 Model Type Evaluation

To study the effect of the novel ensemble ML techniques on TCM performance, as well as the effects of the heterogeneous and homogeneous ensemble configuration types, the average performance metrics across these groups are evaluated and compared to those for the averaged base models. A summary of these results is shown in Table 5.2, and Table 5.3 evaluates the

significance of the 10-fold CV and LOGO-CV score differences between the groups. It is clear that the ensemble ML techniques show strong advantages over individual learners, as the averaged ensemble group scored better than the individual model group in every performance metric measured. A statistical t-test comparing the average individual models' and average ensemble models' 10-fold CV and LOGO-CV scores gives p-values of 1.40e-14 and 1.89e-7 respectively, which confirms the strong statistical significance of these differences. The ensemble group also showed lower score variance than the base model group in every area except for one. These improvements in model performance and prediction stability due to the implementation of advanced ensemble machine learning techniques are consistent with the results from the few existing ensemble ML studies in TCM [6, 13, 14, 48-52]. However, the presented evaluation dives deeper than previous studies by studying a wider range of ensemble and base models, including the ESV and ET models which had not been applied to signal-based TCM before, as well as by evaluating the models' generalizability to data from experiments which they were not trained on.

Table 5.2: Summarized performance results for model groups

		Base ML Models	All Ensembles	Heterogeneous Ensembles	Homogeneous Ensembles
	ML Models Averaged:	DT, SVM, kNN, & ANN	EHV, ESV, Stacked SVM, RF,	EHV, ESV, & Stacked SVM	RF & ET
Experiment Generalizability Study Averages	Average Accuracy (Weighted Recall) Score:	0.8565	0.8756	0.8807	0.8678
	Standard Deviation of Accuracy Score:	0.0209	0.0135	0.0185	0.0060
	95% Confidence Interval of Accuracy Mean:	0.0129	0.0084	0.0114	0.0037
	Macro-Averaged Recall Score:	0.7749	0.7833	0.7719	0.8004
	Weighted Precision Score:	0.8841	0.8986	0.8953	0.9035
	Macro-Averaged Precision Score:	0.8198	0.8539	0.8404	0.8742
	Weighted F1 Score:	0.8408	0.8629	0.8669	0.8570
	Macro-Averaged F1 Score:	0.8198	0.8599	0.8504	0.8742
CV Scores using all Experiments	10-Fold CV Mean Score:	0.9751	0.9823	0.9785	0.9880
	10-Fold CV Score Standard Deviation:	0.0012	0.0009	0.0011	0.0006
	LOGO-CV Mean Score:	0.9122	0.9207	0.9180	0.9248
	LOGO-CV Score Standard Deviation:	0.0032	0.0019	0.0018	0.0021
Tests within 1 Experiment	Average Accuracy for Same Experiment:	0.9746	0.9810	0.9777	0.9858
	Same Experiment Accuracy SD:	0.0025	0.0033	0.0042	0.0021

Table 5.3: Statistical analysis of model group performance differences

	Model 1: Best-Performing Group	Homogeneous Ensemble Models			
Using 20x-Repeated 10-fold CV Scores	Model 2: Comparison Group	Base Models	Ensemble Models	Heterogeneous Models	Homogeneous Models
	t-value:	68.89	25.04	32.79	
	p-value:	3.76E-28	4.01E-17	1.60E-19	
	p-value < 0.05 ?	Yes	Yes	Yes	
	Model 1: Best-Performing Group	Homogeneous Ensemble Models			
Using 20x-Repeated LOGO-CV Scores	Model 2: Comparison Group	Base Models	Ensemble Models	Heterogeneous Models	Homogeneous Models
	t-value:	15.10	7.33	11.68	
	p-value:	1.98E-13	3.22E-07	1.19E-10	
	p-value < 0.05 ?	Yes	Yes	Yes	

Within the group of ensemble ML models, the performances of the heterogeneous and homogeneous configurations were also compared. It was found that, similar to the stacked SVM results, the heterogeneous ensemble group performed the best in the accuracy and weighted F1 scores, and the homogeneous group did better in all of the remaining metrics. The homogeneous ensemble group also generally achieved higher prediction stability than the heterogeneous group. These results are likely due to the RF and ET models' much larger numbers of base models to draw information from (100, compared to 4), as well as their high generalization ability and effectiveness for unbalanced datasets [14, 54].

In summary, across the diverse range of performance metrics and model generalization studies presented, the ensemble methods showed clear performance advantages over the individual models, the homogeneous ensemble group performed better than the heterogeneous group, and the extra-trees algorithm was shown to be the best-performing model for this application.

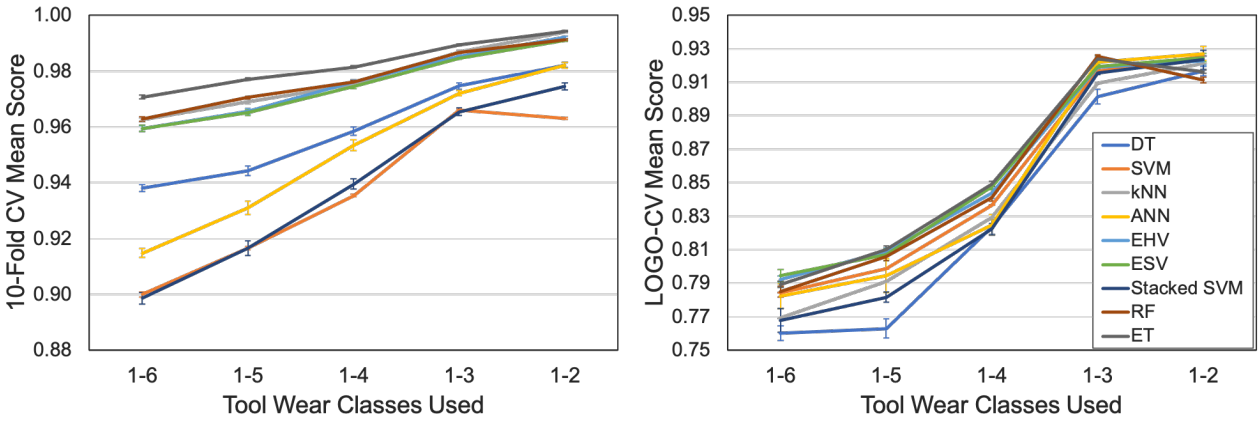
5.1.2 Effect of Classification Resolution

The choice of a TCM system's tool wear classification resolution is important, as it balances prediction accuracy with output precision. The choice may also depend on project-

specific variables such as the wear measurement or range of measurements at which a tool should be changed to ensure product quality, or the amount of time to be allotted between a wear notification and when an operator can be expected to change the tool. To study the effect of tool wear classification precision on the different models' performance, the ML samples were re-labeled according to Table 5.4. The resulting model 10-fold CV and LOGO-CV scores are shown in Figure 5.1.

Table 5.4: Tool wear classification resolutions studied

Wear Resolution:	1-6	1-5	1-4	1-3	1-2
Wear Level	VB Range (mm)				
1	0.00-0.05	0.00-0.10	0.00-0.15	0.00-0.10	0.00-0.15
2	0.05-0.10	0.10-0.15	0.15-0.20	0.10-0.20	0.15-0.30
3	0.10-0.15	0.15-0.20	0.20-0.25	0.20-0.30	
4	0.15-0.20	0.20-0.25	0.25-0.30		
5	0.20-0.25	0.25-0.30			
6	0.25-0.30				



(a) **(b)**

Figure 5.1: Effect of classification resolution on model (a) 10-fold CV, and (b) LOGO-CV scores

These results show that as the number of tool wear classes is reduced, the models' prediction accuracy generally increases. The largest increase in both 10-fold CV and LOGO-CV scores

occurs between wear resolutions 1-4 and 1-3, and the lowest difference is from 1-3 to 1-2. Based on this, as well as the high scores across both metrics and all model types, a classification resolution of three wear levels from 0.0 – 0.3 mm flank wear is shown to be a good balance of model prediction resolution and precision for TCM. The ET model's high scores across all of the wear resolution options shows a high level of model flexibility, and supports its selection for TCM applications in Section 5.1.1.

However, if a higher level of wear resolution is desired, this can be achieved satisfactorily by using the higher classification ranges presented. For example, the confusion matrices for four models are shown in Figure 5.2 after they were each trained on data from Experiments 1-4 and tested on data from Experiments 5-8 using a classification resolution of 6 wear levels.

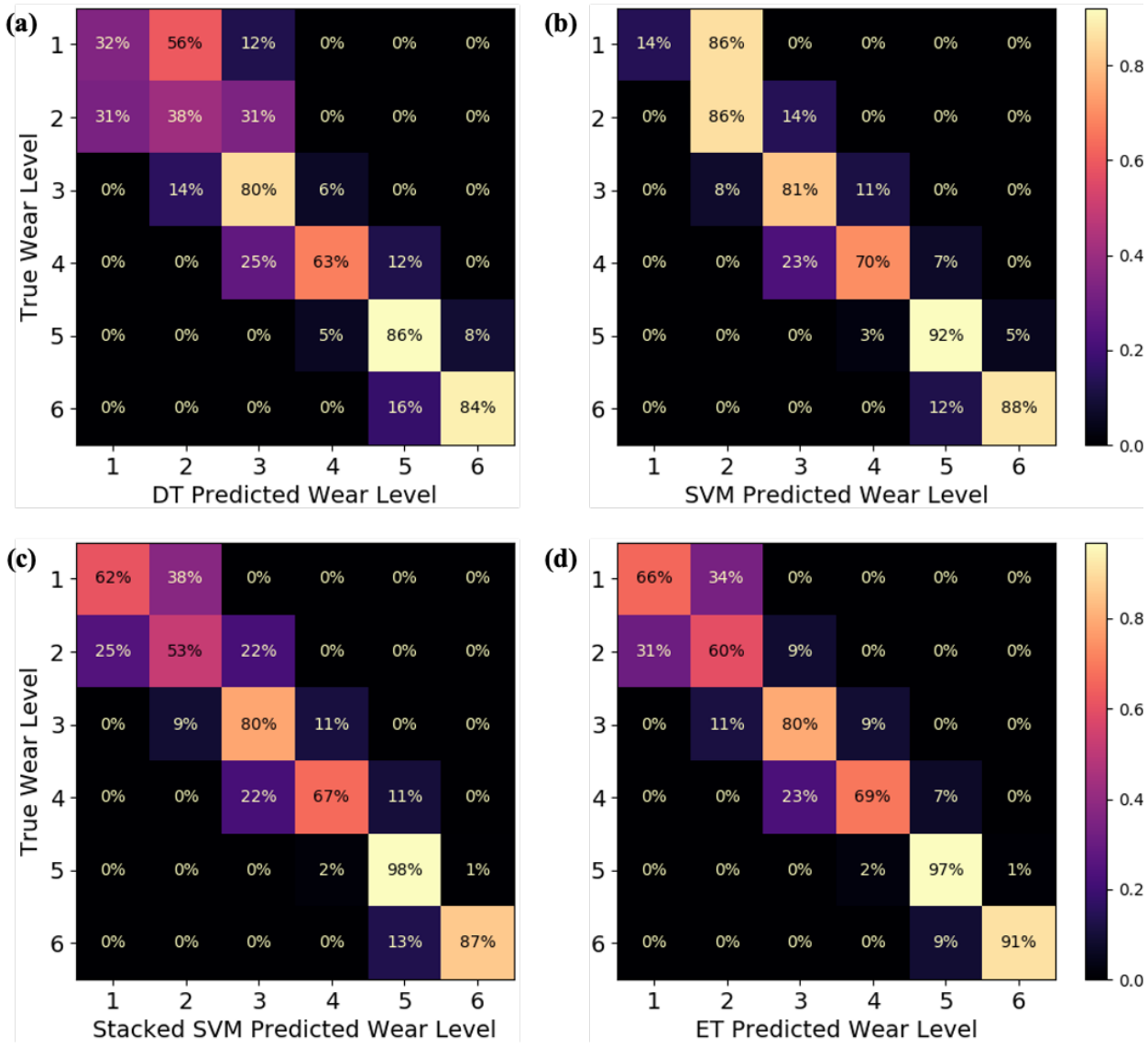


Figure 5.2: Confusion matrices for (a) DT, (b) SVM, (c) Stacked SVM, and (d) ET models when trained on data from Experiments 1-4 and tested on Experiment 5-8 using a classification resolution of 6 levels

Impressively, even for all nine models and such a high level of resolution, almost no samples were misclassified by more than one wear level—this only occurred for the DT and a small number of samples between wear levels 1 and 3 as shown. For the classification resolution of 6 levels, the accuracy scores ranged from 79.0% for the DT, to 84.0% for the EHV, with the ensemble models generally achieving higher accuracy scores than the individual models. In agreement with the

results discussed in Section 4.4.2, the predictions for the lowest numbered wear classes, which had the least true samples available, were improved the most by the application of ensemble ML techniques.

Through the presented TCM system configuration studies presented here and in Section 4.3, it is found that a system's classification performance can be optimized through the use of the SVM-RFECV feature selection method, a homogeneous ensemble ML analysis technique such as the especially strong extra-trees ensemble model, and a wear classification resolution of 3 levels. In addition, 6 classes of wear resolution also gives impressive results if a higher level of process visibility is desired.

5.2 Machining Condition Generalizability Study

To build an understanding of TCM system generalizability to new machining conditions, each ML model is evaluated on a variety of different experiments and cutting parameter sets. As shown in Table A.2, ML runs are set up to test model generalizability to data from new experiments which used the same cutting conditions as the training set, which used a different feed rate or spindle speed compared to the training set, or which used both a different feed rate and a different spindle speed compared to the training set. Models were also trained on various experiment combinations in order to assess how increased machining condition variability in training sets affects model generalization performance. None of these ML runs included data from any one experiment in both the training and testing sets, ensuring the validity of the results for practical TCM applications. However, the results of four runs in which models are trained and tested on different subsets of the same experiment's data are shown in Table A.3 for comparison.

Figure 5.3 shows each model's average TCM performance across various cutting condition or experiment changes. Each ML test is run 20 times and the accuracy scores are averaged across these runs and across the other ML tests with the same process changes. The error bars give the 95% confidence intervals of these mean values.

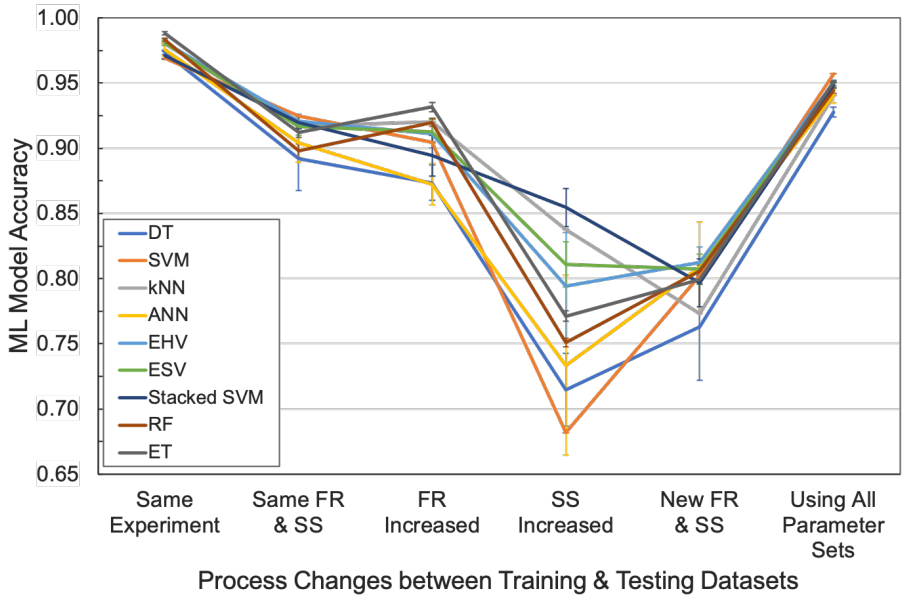


Figure 5.3: Effects of machining condition changes on model performance

As expected, it is observed that the models all achieved their highest classification accuracy scores when they were trained and tested on different subsets of data from the same experiment. While commonly applied in TCM literature when more data is not available, this case demonstrates the lowest possible level of machining condition variability and is a problematic measure of TCM model effectiveness due to its low applicability to real-world situations when classification systems must be pre-trained [16, 169].

The next level of machining condition variability is assessed by training the models on data from one experiment (i.e. 1A), and testing them on a different experiment which used the same

cutting parameters (i.e. 8A). Due to the high levels of variability in tool wear rates, patterns, and process signals between individual experiments due to the large number of effectively random variables involved in metal milling processes [19, 20], the models' accuracies all dropped significantly compared to the previous test.

While individual ML model results differ more for the studies in which cutting parameters are changed, it is observed that overall, changes in feed rate show little effect on model performance, while changes in spindle speed caused dramatic performance changes. This correlates to previous literature in which machining tool wear rates have been found to depend much more on cutting speed settings than on the feed rates or chip loads used [2, 5, 36, 82-84]. For this study the impacts of the cutting speed on model performance are effectively the same as those of the spindle speed, due to the constant tool diameter.

The final ML test set shown in Figure 5.3 assesses how well the models perform when they are trained on Experiments 1-4 and tested on Experiments 5-8. As each of these experiment groups includes one experiment from each of the four cutting parameter sets, the model performances from this analysis can be compared to those in which the models were trained on one experiment and tested on a new experiment which used the same cutting parameters. This comparison shows that an increase in machining condition variability in the training set can significantly improve model generalizability to data from new experiments, which is an important quality for practical TCM systems [19, 28]. This result is found to be consistent across all nine machine learning models applied in this study.

Figure 5.4 shows a more detailed view of how different experiments' chip load values can affect TCM model performance. As the chip load parameter is based on both the spindle speed and

the feed rate in milling processes, as described in Equation 2.2, the results from four chip load values can be evaluated for this study.

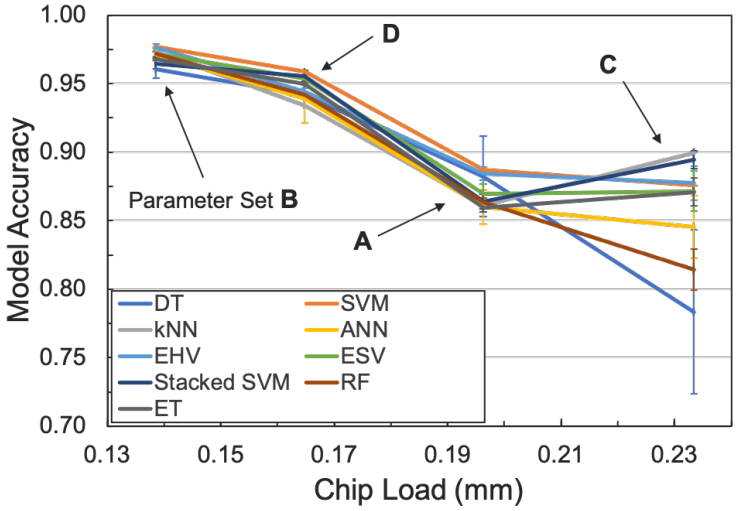


Figure 5.4: Model performance at different chip loads when trained and tested on different experiments of the same cutting parameter set

It is observed that an increase in the chip load used for a set of experiments results in fairly consistent decreases in model generalizability from one experiment to another. The differences in model performance can be large—the DT model, for example, decreased in mean accuracy from 96.1% to 78.3% across the range of chip loads studied, while even the least-affected models' accuracies decreased by at least 7% and had wide gaps between their 95% confidence intervals. As a higher chip load is correlated to increased tool wear rates, although to a lesser extent than cutting speed is [2, 5, 82-84], the results of the ML analyses summarized in both Figure 5.3 and Figure 5.4 suggest that TCM model performance and generalizability may be inversely correlated with milling tool wear rates. A comparison of these results with the experiments' sample sizes, listed in Table 4.1, does not find any sampling patterns which would cause this effect.

Through this analysis of TCM system generalizability to new machining conditions and individual experiments, the importance of evaluating models on unseen experiments, the advantages of training models on data from various machining conditions, and a possible inverse relationship between tool wear rates and model performance are identified.

5.3 Environmental Noise Generalizability Improvement Study

Finally, the effect of the noisy training technique on TCM model generalizability is studied for the first time. To do this, the nine ML model types are each trained on various levels and combinations of white Gaussian noise-augmented datasets, and their generalization ability to data from an unseen experiment is evaluated. Each model type, several values for the percentage of the training dataset which is made up of noise-augmented data, and 4 injected noise levels are investigated. Two sets of experimental data are noise-augmented for this study. First, the original data from Experiment 3C is injected with artificial noise and the ML models trained on combinations of noisy and original data are tested on the Experiment 6C original data. Second, the original data from Experiment 2B is injected with artificial noise and these trained ML models are tested on the Experiment 7B original data. The levels of AWGN added to the 3C and 2B datasets are listed in Table 3.6. The artificial noise at each level is added to duplicates of the entire experiment's dataset each time, effectively generating large amounts of new data for training the models. Two tool wear classification resolutions are also investigated: 3 wear levels and 6 wear levels. The training and testing set layouts for this noisy training study are detailed in Table 5.5. Each test is repeated 20 times for each ML model, and the resulting mean scores and 95% confidence intervals of the means are shown below.

Table 5.5: Noisy training study layout

				Training Dataset					Testing Dataset
Test Number	Wear Resolution	Analysis	Percentage of Training Samples which are Noise-augmented	L0 (no added noise)	L1	L2	L3	L4	L0 (no added noise)
1	3 Wear Levels	Percentage of Training Set which is Augmented	0%	3C					6C
2			50%	3C	3C				6C
3			67%	3C	3C	3C			6C
4			75%	3C	3C	3C	3C		6C
5			80%	3C	3C	3C	3C	3C	6C
1		Injected Noise SNR Level	0%	3C					6C
2			50%	3C	3C				6C
6			50%	3C		3C			6C
7			50%	3C			3C		6C
8			50%	3C				3C	6C
9	6 Wear Levels	Percentage of Training Set which is Augmented	0%	3C					6C
10			50%	3C	3C				6C
11			67%	3C	3C	3C			6C
12			75%	3C	3C	3C	3C		6C
13			80%	3C	3C	3C	3C	3C	6C
9		Injected Noise SNR Level	0%	3C					6C
10			50%	3C	3C				6C
14			50%	3C		3C			6C
15			50%	3C			3C		6C
16			50%	3C				3C	6C
17	6 Wear Levels	Percentage of Training Set which is Augmented	0%	2B					7B
18			50%	2B	2B				7B
19			67%	2B	2B	2B			7B
20			75%	2B	2B	2B	2B		7B
21			80%	2B	2B	2B	2B	2B	7B
17		Injected Noise SNR Level	0%	2B					7B
18			50%	2B	2B				7B
22			50%	2B		2B			7B
23			50%	2B			2B		7B
24			50%	2B				2B	7B

Based on promising results from the neural computing and speech recognition communities surrounding the use of noisy training for neural networks [33-35, 140-144], it is expected that the ANN model, at least, will see reduced overfitting and improved generalizability for low levels of AWGN included in the training sets. However, noise levels which are too high have previously focused the models' training on the noise patterns too much and hurt their performance overall [33]. Therefore, it is important to determine the correct groups and levels of noise to be added, based on the specific data and application.

The results of the 3C noisy training study using a wear classification resolution of 3 levels (tests 1-8 in Table 5.5) are detailed in Table A.4. The error reduction percentages are also shown for all nine models in Figure 5.5. In addition, the results are shown more clearly for the average effects across models, as well as the most significantly affected models, in Figure 5.6. In these figures, the error reduction percentages are calculated based on the models' prediction accuracies when they are first trained on the original, unaltered Experiment 3C data and tested on the original 6C data (test 1).

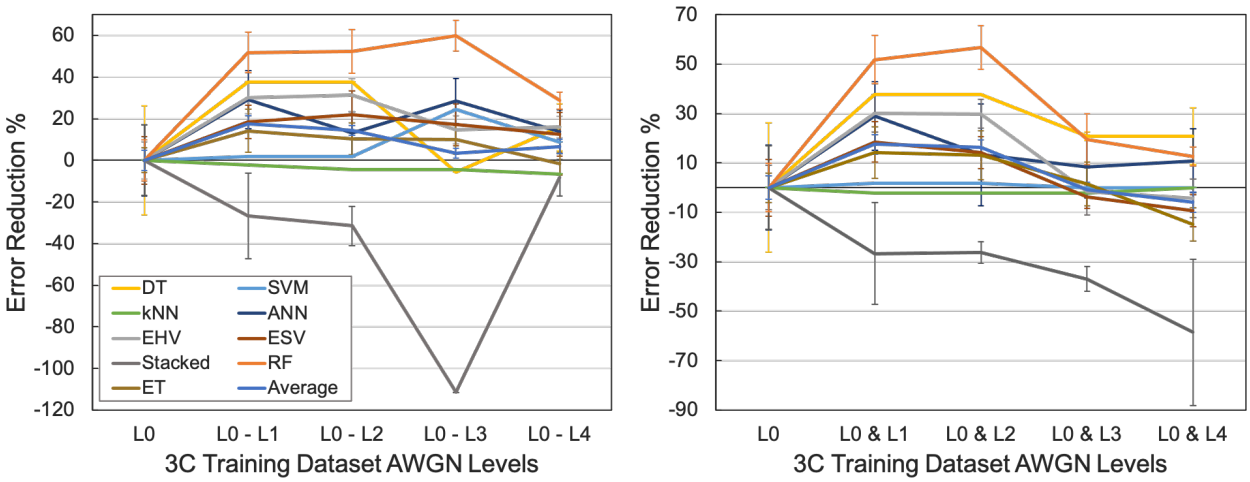


Figure 5.5: Wear 1-3 classification error reduction with noisy training technique, using (a) various combinations of noisy datasets added to the 3C original data, and (b) various AWGN SNR levels for one dataset added to the 3C original data

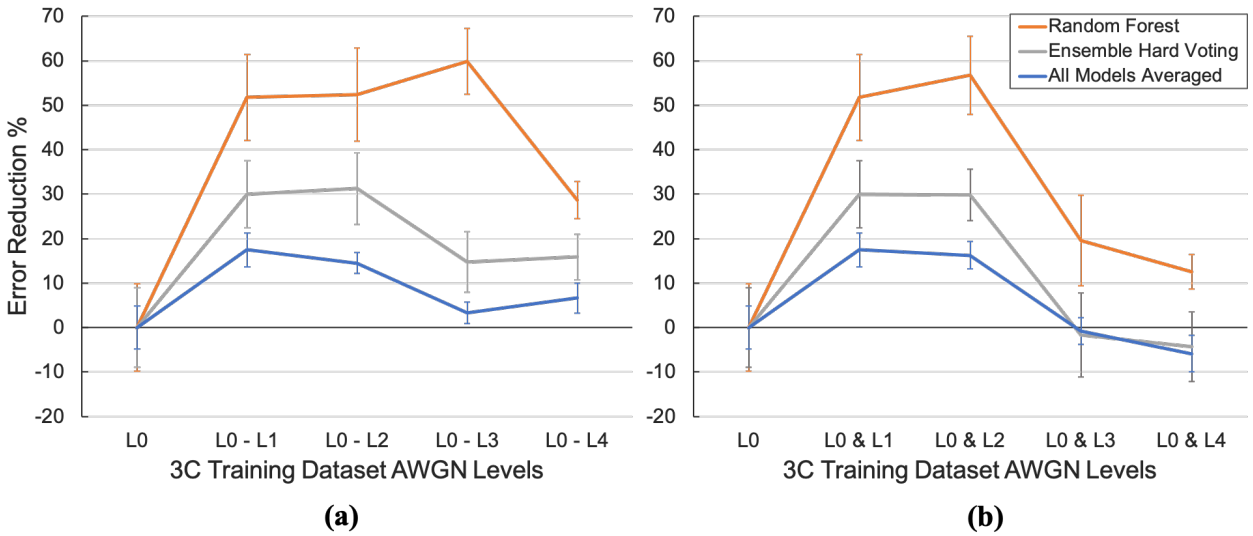


Figure 5.6: Summary of wear 1-3 classification error reduction with noisy training, using (a) various combinations of noisy datasets added to the 3C original data, and (b) various AWGN SNR levels for one dataset added to the 3C original data

From Figure 5.5 it is observed that 7 out of the 9 ML models saw fairly consistent accuracy improvements in the first few categories studied, especially for the combination of L0 and L1 (the original and SNR25 noise) datasets in the training set. These results are found for both the various combinations of noisy datasets used to train the models, and the different levels of noise injected into just one additional dataset. After these model generalizability improvements at the lowest levels of noise, the models' accuracies drop for higher noise levels, which is expected based on past neural network studies [33].

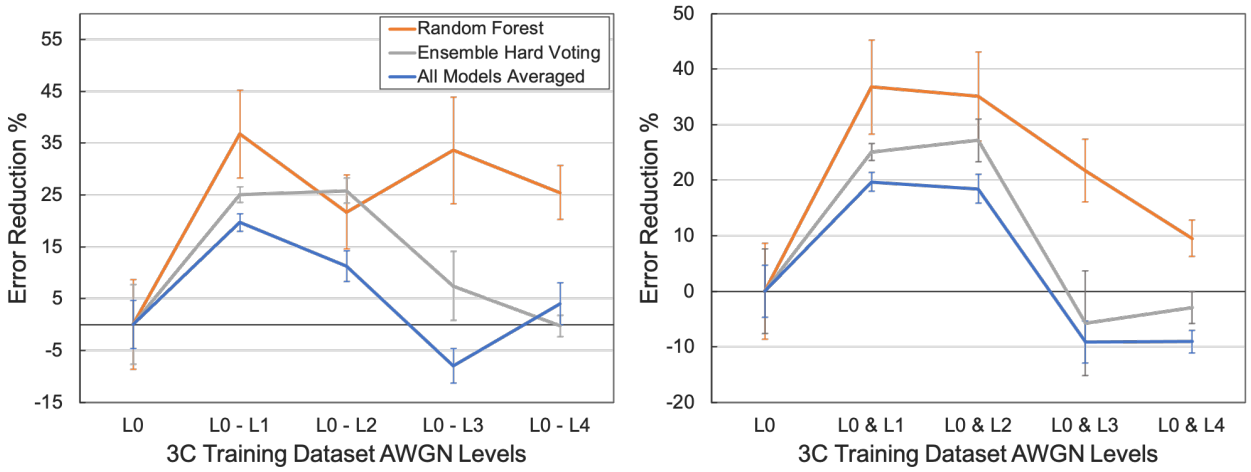
The t-tests summarized in Table 5.6 assess the statistical significance of the models' performance differences when they are trained on the L0 and L1 datasets (test 2), compared to when no noise-augmented datasets are used (test 1). Of the seven improved models, six of the results are shown to be statistically significant using a p-value of 0.05. Only the decision tree's accuracy increase is found to be insignificant, as its original model variance is too high to prove that the recorded performance improvement does not occur randomly.

Table 5.6: Statistical analysis of noisy training generalization effects

Comparison 1:	Trained on only **3C L0** (original data) & Tested on 6C original data								
Comparison 2:	Trained on **3C L0 + L1** data & Tested on 6C original data								
ML Model	DT	SVM	kNN	ANN	EHV	ESV	Stacked SVM	RF	ET
t-value:	-1.86	-3.67E+13	2.45E+13	-2.11	-5.36	-2.57	2.47	-4.22	-2.47
p-value:	0.074	0.00E+00	0.00E+00	0.049	4.24E-05	0.019	0.027	5.20E-04	0.024
p-value < 0.05 ?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

The stacked generalization SVM ensemble method is the only technique which is shown to be severely negatively affected by the noise augmentation in its training datasets. The kNN model is only slightly affected by the noisy training, which is consistent with Xing et al.'s findings that kNN models are already highly robust to noisy samples and therefore do not respond to the noisy training technique [152].

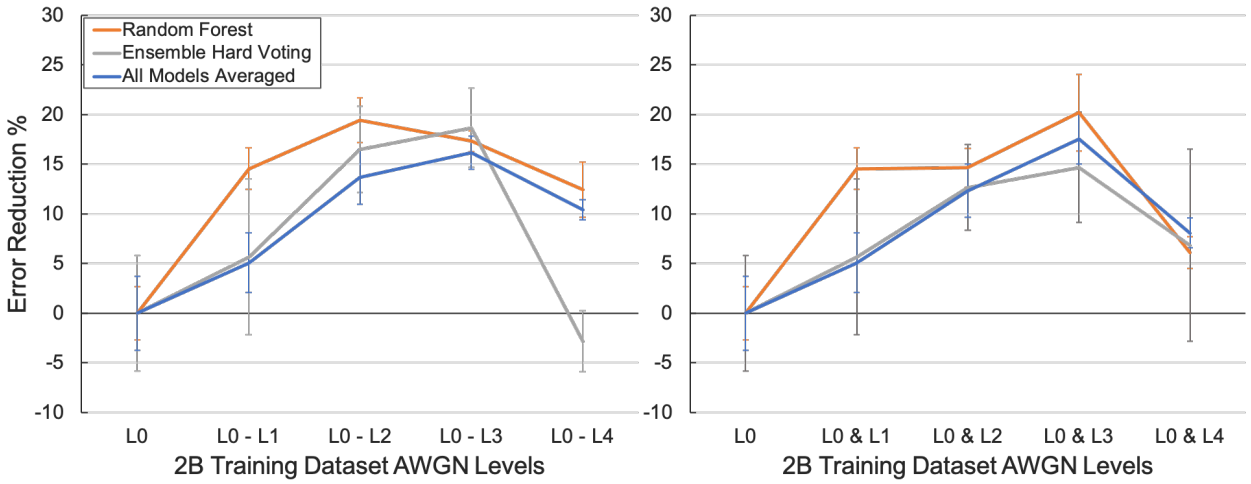
In order to evaluate the validity and robustness of these model generalization improvements due to the noisy training technique, the tests are repeated for Experiments 3C and 6C, as well as for 2B and 7B, both using wear classification resolutions of 6 wear levels. The summarized results of these studies are shown in Figure 5.7 and Figure 5.8 respectively.



(a) **(b)**

Figure 5.7: Summary of wear 1-6 classification error reduction with noisy training, using (a) various combinations of noisy datasets added to the 3C original data, and (b) various AWGN SNR levels for one dataset added to the 3C original data

The results for the 3C vs. 6C noisy training study using a wear resolution of 6 levels (tests 9-16) confirm the conclusions drawn for 3 wear levels, with optimal model generalizability to the new experiment's data being seen when the models are trained on the L0 and L1 datasets together.



(a) **(b)**

Figure 5.8: Summary of wear 1-6 classification error reduction with noisy training technique, using (a) various combinations of noisy datasets added to the 2B original data, and (b) various AWGN levels for one dataset added to the 2B original data

The third noisy training study, in which the new 2B and 7B experimental data is used (tests 17-24), shows similar results and large reductions in original prediction error through the use of the L0 and L3 datasets for training the models (test 23). After this optimal level of noise, as observed for the parameter set C studies, the prediction accuracies decrease.

Overall, the results from the three noisy training analyses are fairly consistent in that they all enable significant increases in model generalizability to data from a new and different dataset for certain injected noise SNR ranges. While an SNR value of 25 was found to be optimal for the 3C vs. 6C studies using wear resolutions of either 3 or 6, an SNR value of 15 performed better for the 2B vs. 7B study. Using these training datasets, the average tool wear classification errors are reduced by 17.5%, 19.7%, and 17.5% compared to when no noise augmentation is used. Based on these results, the optimal level of injected noise is dependent on the specific datasets used within TCM, but is independent of the classification resolution applied.

In addition, for all three studies, the highest error reduction was achieved with the addition of only one noisy dataset and therefore a percentage of noisy training data of 50%, suggesting that adding larger numbers of noisy samples to the training dataset past this level decreases the model's ability to focus on important TCM process information extracted from the original data. These results demonstrate not only that the noisy training technique can be applied effectively to other ML models besides neural networks, such as DT, SVM, EHV, ESV, stacked SVM, RF, and ET, but also that this could be a promising method of significantly reducing TCM classification error and improving generalizability, especially in situations where limited original data is available for model training.

CHAPTER 6: CONCLUSION

This study aimed to help fill the research gaps surrounding tool condition monitoring system generalizability, ensemble machine learning techniques, and practical sensor fusion. To do this, nine ML models were investigated, including five ensemble ML methods, based on indirect signals from practical external and internal machine sensors. Several feature selection and classification resolution options were also evaluated for this application. Finally, using original experimental data from milling tool life experiments conducted across various machining conditions, as well as simulated data using various levels of injected noise, the models' generalizability is assessed and methods of improving it are evaluated.

6.1 Contributions

By addressing the significant research gaps, this study presents several contributions to the TCM research body. These include:

- Several ensemble ML methods were applied to this field for the first time, or were investigated for TCM applications outside the few which had been addressed previously. Overall, the ensemble models performed significantly better than the individual models in every performance metric, and the homogeneous ensemble models generally performed better than the heterogeneous models.
- The extra-trees ensemble method, which had never before been studied for signal-based TCM, performed the best out of all nine ML models studied. With its high levels of accuracy, generalizability to new conditions, effectiveness for unbalanced datasets, and classification speed, it is a viable option for in-process tool condition monitoring.

- Five signal feature selection methods were studied across multiple ML models and performance metrics. The SVM-based recursive feature elimination method was found to achieve the best results for TCM.
- LOGO-CV scores were found to be a better measure of model generalizability and performance for industrial use than k-fold CV scores, even when both use data from various machining conditions. The extra-trees ensemble ML model achieved mean LOGO-CV and 10-fold CV accuracy scores of 92.4% and 99.0%, respectively.
- A method of developing and evaluating TCM systems in a way which is more realistic for predicting commercial results is presented, and its use may help future models be compared more easily.
- Through the analysis of ML model generalizability across changes in experiments' spindle speed, feed rate, and chip load, an inverse relationship between a tool's wear rate and classification performance is identified. Specifically, changes in spindle speed had a large effect on model performance, the chip load had some effect, and the feed rate had little effect. It is also found that increasing machining condition variability in the training dataset can improve model generalizability.
- The noisy training method was found to be successful at significantly improving model generalizability for the SVM, ANN, EHV, ESV, RF, and ET models. Averaged across all nine models, the mean error reduction at the identified optimal SNR values ranged from 17.5% to 19.7%. This technique had not been evaluated for generalizability improvement in TCM previously, and addresses the industry's need for more generalizable models.

6.2 Assumptions and Limitations

Several assumptions have been made during the completion of this study, and its conclusions are not without limitations. First, it is assumed that the 1-second segments of data along each machining pass do not differ significantly for reasons unrelated to tool wear, and that therefore the machining pass signals' segmentation into 1-second segments for analysis is appropriate. The evaluated models do not take process parameters such as the position of the tool in the material or the distance between the tool and the microphone into account, and assume that changes in the signals' extracted features are a result of a tool's wear level. While this assumption is common in existing literature [2, 22, 23, 29, 30, 48] and allows more process information to be extracted by ML algorithms, it also causes the TCM system to be more reliant on the machining paths and conditions remaining relatively constant during each pass. However, the generalizability improvement methods studied in this research aim to reduce this reliance.

Similarly, as the feature extraction methods calculate only one feature value from the 1-second segments of each process signal, it is assumed that the signals and their features do not significantly change over the span of one segment. If a signal does significantly change mid-segment due to process condition changes, environmental changes, or tool chipping, this may result in an inaccurate wear measurement prediction.

A few other limitations result from the fact that only eight tool life experiments were run, using four machining conditions. While the spindle speed and feed rate were varied between two levels each, the tool, the toolpath, the machine, and the other process parameters were kept constant. Due to this, the generalizability of the models to other parameter changes or toolpaths cannot be assessed without more experimental data.

In addition, for the noisy training analysis, the conclusions are limited by the number of noise SNR levels which were studied. While the specific levels were chosen based on previous related studies [33, 157] and the results appear to show the model generalizability improvements at low noise levels and decreases at high noise levels which were expected, the analysis of more noise levels would allow the optimal SNR values and any relevant model performance patterns to be more accurately identified.

6.3 Future Work

While this study helps fill several significant research gaps in TCM, it also presents new opportunities for further research. While the extra-trees ML model showed the best tool wear classification performance for this combination of process signals, it would be beneficial to evaluate it for other sensor fusion combinations to better understand its advantages for TCM. In addition, as the heterogeneous ensemble models studied in this research were always based on one group of four base models, a study in which several heterogeneous ensemble models were studied based on a variety of different base model combinations would better evaluate any possible advantages of heterogeneous ensembles over homogeneous ensembles for TCM.

In addition, a larger number of tool life experiments conducted across a wider range of machining parameters would allow the effects of spindle speed, cutting speed, feed rate, and chip load on tool wear classification performance to be further evaluated, and for the identified potential inverse correlation between wear rate and model performance to be further examined. The effects of other machining parameters on model performance could also be assessed. An analysis of the noisy training technique using a larger number of SNR values and more original data would also help identify the capabilities and limitations of this effect. As different optimal training noise levels

were identified based on the datasets applied in this study, any potential methods of selecting or automatically calculating the optimal noise levels to be used without conducting these extensive experiments would be particularly beneficial. Through the ensemble model configuration and generalizability advancements presented in this study, as well as to be seen in future works like these, the TCM roadblock to the implementation of smart factories could be overcome.

APPENDIX A

Table A.1: Feature selection results

	Feature Selection Method:			
	RFE-CV using SVM	RFE-CV using RF	Mutual Information	Correlation Coefficient
Sound Features Selected:	TD mean TD standard deviation TD skewness TD median FD amplitude sum 0-9kHz		TD mean TD standard deviation TD RMS FD amplitude sum 0-500Hz FD amplitude sum 0-9kHz	
Spindle Power Features Selected:	TD mean TD RMS	TD mean	TD mean TD median TD standard deviation TD maximum TD RMS TD range FD maximum amplitude FD peak at spindle frequency FD amplitude sum 0-83Hz	TD mean TD RMS FD maximum amplitude FD peak at spindle
X-Axis Load Features Selected:	TD standard deviation TD skewness		TD standard deviation TD maximum FD frequency standard deviation TD range	TD standard deviation
Y-Axis Load Features Selected:	TD mean TD maximum TD clearance factor		TD mean TD median TD standard deviation TD maximum TD RMS TD range FD maximum amplitude FD peak at spindle frequency FD amplitude sum 0-83Hz	TD mean TD standard deviation TD maximum TD RMS TD range FD maximum amplitude FD peak at spindle frequency
Z-Axis Load Features Selected:	TD mean TD median	TD standard deviation TD RMS	TD mean TD median TD standard deviation TD maximum TD RMS TD range FD maximum amplitude FD peak at spindle frequency FD amplitude sum 0-83Hz	TD mean TD standard deviation TD maximum TD RMS TD range FD maximum amplitude FD peak at spindle frequency FD amplitude sum 0-83Hz

Table A.2: Analysis of model transferability to new machining conditions

		Same feed rate & spindle speed				Feed rate increased		Spindle speed increased		New feed rate & spindle speed		Both using all cutting parameter sets	
	Training Data:	1A	2B	3C	4D	2B, 7B	4D, 5D	3C, 6C	4D, 5D	1A, 8A	2B, 7B	1A, 2B, 3C, & 4D	Score
Model:	Testing Data:	8A	7B	6C	5D	1A, 8A	3C, 6C	1A, 8A	2B, 7B	4D, 5D	3C, 6C	5D, 6C, 7B, & 8A	Average:
DT	Average Accuracy (Weighted Recall) Score:	0.8823	0.9608	0.7834	0.9417	0.9204	0.8266	0.7242	0.7050	0.6865	0.8392	0.9277	0.8362
	Standard Deviation of Accuracy Score:	0.0473	0.0105	0.0966	0.0037	0.0339	0.0100	0.0809	0.0094	0.1244	0.0074	0.0063	0.0391
	95% Confidence Interval of Accuracy Mean:	0.0293	0.0065	0.0598	0.0023	0.0210	0.0062	0.0501	0.0058	0.0771	0.0046	0.0039	0.0242
	Macro-Averaged Recall Score:	0.6719	0.9624	0.5577	0.9621	0.8020	0.8081	0.7455	0.7751	0.6395	0.6576	0.8811	0.7694
	Weighted Precision Score:	0.8854	0.9706	0.7807	0.9617	0.9344	0.8522	0.8972	0.7918	0.7244	0.8430	0.9277	0.8699
	Macro-Averaged Precision Score:	0.9173	0.9765	0.5655	0.8650	0.9262	0.7546	0.8139	0.7500	0.6216	0.8485	0.8915	0.8119
	Weighted F1 Score:	0.8636	0.9682	0.7437	0.9489	0.9009	0.8370	0.7281	0.6993	0.6630	0.8174	0.9277	0.8271
	Macro-Averaged F1 Score:	0.9173	0.9765	0.5655	0.8650	0.9262	0.7546	0.8139	0.7500	0.6216	0.8485	0.8915	0.8119
SVM	Average Accuracy (Weighted Recall) Score:	0.8872	0.9768	0.8758	0.9589	0.9031	0.9064	0.8177	0.5460	0.8133	0.7911	0.9571	0.8576
	Standard Deviation of Accuracy Score:	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	95% Confidence Interval of Accuracy Mean:	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Macro-Averaged Recall Score:	0.7955	0.9723	0.6293	0.9712	0.6626	0.6607	0.8505	0.6608	0.8114	0.5968	0.9479	0.7781
	Weighted Precision Score:	0.8936	0.9773	0.8273	0.9677	0.9150	0.9153	0.9869	0.7791	0.8443	0.8148	0.9578	0.8981
	Macro-Averaged Precision Score:	0.9159	0.9820	0.5981	0.8905	0.9126	0.9426	0.9681	0.8288	0.7858	0.8625	0.9353	0.8748
	Weighted F1 Score:	0.8818	0.9769	0.8444	0.9610	0.8755	0.8713	0.7901	0.4002	0.8082	0.7604	0.9571	0.8297
	Macro-Averaged F1 Score:	0.9159	0.9820	0.5981	0.8905	0.9126	0.9426	0.9681	0.8288	0.7858	0.8625	0.9353	0.8748
kNN	Average Accuracy (Weighted Recall) Score:	0.8609	0.9735	0.8998	0.9340	0.9425	0.8975	0.7576	0.9179	0.7959	0.7507	0.9405	0.8792
	Standard Deviation of Accuracy Score:	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	95% Confidence Interval of Accuracy Mean:	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Macro-Averaged Recall Score:	0.7466	0.9466	0.6538	0.9460	0.6798	0.6888	0.7341	0.8884	0.6917	0.5528	0.9170	0.7678
	Weighted Precision Score:	0.8756	0.9704	0.8521	0.9456	0.9020	0.8697	0.9043	0.9328	0.8003	0.7012	0.9406	0.8813
	Macro-Averaged Precision Score:	0.9177	0.9764	0.6233	0.8587	0.6276	0.7240	0.7942	0.9125	0.6954	0.5028	0.9127	0.7768
	Weighted F1 Score:	0.8584	0.9683	0.8685	0.9367	0.9191	0.8796	0.7660	0.9190	0.7818	0.7079	0.9403	0.8678
	Macro-Averaged F1 Score:	0.9177	0.9764	0.6233	0.8587	0.6276	0.7240	0.7942	0.9125	0.6954	0.5028	0.9127	0.7768
ANN	Average Accuracy (Weighted Recall) Score:	0.8598	0.9714	0.8455	0.9390	0.8337	0.9106	0.7105	0.7563	0.7580	0.8600	0.9407	0.8532
	Standard Deviation of Accuracy Score:	0.0201	0.0073	0.0370	0.0290	0.0347	0.0164	0.1136	0.1091	0.0602	0.0511	0.0097	0.0444
	95% Confidence Interval of Accuracy Mean:	0.0125	0.0045	0.0230	0.0180	0.0215	0.0102	0.0704	0.0676	0.0373	0.0317	0.0060	0.0275
	Macro-Averaged Recall Score:	0.7501	0.9649	0.6126	0.9669	0.6699	0.7680	0.6807	0.7963	0.6685	0.8121	0.9381	0.7844
	Weighted Precision Score:	0.8741	0.9732	0.8197	0.9636	0.8293	0.9022	0.9345	0.8209	0.8136	0.8783	0.9488	0.8871
	Macro-Averaged Precision Score:	0.8981	0.9712	0.6421	0.8844	0.6905	0.8221	0.8083	0.8627	0.7096	0.7698	0.9121	0.8155
	Weighted F1 Score:	0.8542	0.9721	0.8102	0.9555	0.8219	0.8802	0.6578	0.7521	0.7226	0.8517	0.9468	0.8386
	Macro-Averaged F1 Score:	0.8981	0.9712	0.6421	0.8844	0.6905	0.8221	0.8083	0.8627	0.7096	0.7698	0.9121	0.8155
EHV	Average Accuracy (Weighted Recall) Score:	0.8845	0.9763	0.8775	0.9442	0.9104	0.9119	0.7973	0.7909	0.7913	0.8331	0.9502	0.8789
	Standard Deviation of Accuracy Score:	0.0079	0.0043	0.0200	0.0041	0.0317	0.0040	0.0268	0.1067	0.0219	0.0166	0.0033	0.0225
	95% Confidence Interval of Accuracy Mean:	0.0049	0.0027	0.0124	0.0025	0.0196	0.0025	0.0166	0.0661	0.0136	0.0103	0.0021	0.0139
	Macro-Averaged Recall Score:	0.7776	0.9557	0.6224	0.9694	0.7124	0.6834	0.7600	0.8100	0.6839	0.6265	0.9113	0.7739
	Weighted Precision Score:	0.8809	0.9730	0.8220	0.9696	0.9145	0.8653	0.9675	0.8463	0.8102	0.8185	0.9422	0.8918
	Macro-Averaged Precision Score:	0.9234	0.9785	0.5944	0.8773	0.8738	0.7201	0.9070	0.8407	0.7043	0.7614	0.9134	0.8268
	Weighted F1 Score:	0.8615	0.9714	0.8355	0.9591	0.8958	0.8751	0.8006	0.7857	0.7732	0.7992	0.9422	0.8636
	Macro-Averaged F1 Score:	0.9234	0.9785	0.5944	0.8773	0.8738	0.7201	0.9072	0.8407	0.7043	0.7614	0.9134	0.8268
ESV	Average Accuracy (Weighted Recall) Score:	0.8698	0.9719	0.8716	0.9538	0.9161	0.9097	0.8054	0.8160	0.7904	0.8242	0.9486	0.8798
	Standard Deviation of Accuracy Score:	0.0110	0.0059	0.0238	0.0043	0.0197	0.0089	0.0167	0.0393	0.0221	0.0155	0.0023	0.0154
	95% Confidence Interval of Accuracy Mean:	0.0068	0.0037	0.0147	0.0026	0.0122	0.0055	0.0104	0.0244	0.0137	0.0096	0.0014	0.0096
	Macro-Averaged Recall Score:	0.7136	0.9532	0.6120	0.9721	0.6709	0.6896	0.7435	0.8404	0.7019	0.6337	0.9155	0.7679
	Weighted Precision Score:	0.8737	0.9719	0.8138	0.9725	0.9143	0.8679	0.9731	0.8559	0.8200	0.8314	0.9452	0.8945
	Macro-Averaged Precision Score:	0.9141	0.9776	0.5887	0.8822	0.8623	0.7316	0.9498	0.8512	0.7279	0.8411	0.9169	0.8403
	Weighted F1 Score:	0.8566	0.9701	0.8220	0.9627	0.8953	0.8763	0.8047	0.7980	0.7801	0.7942	0.9451	0.8641
	Macro-Averaged F1 Score:	0.9141	0.9776	0.5887	0.8822	0.8623	0.7316	0.9498	0.8512	0.7279	0.8411	0.9169	0.8403
Stacked SVM	Average Accuracy (Weighted Recall) Score:	0.8641	0.9647	0.8943	0.9556	0.8926	0.8961	0.7889	0.9205	0.7566	0.8367	0.9483	0.8835
	Standard Deviation of Accuracy Score:	0.0086	0.0065	0.0106	0.0063	0.0406	0.0102	0.0322	0.0149	0.0353	0.0241	0.0035	0.0175
	95% Confidence Interval of Accuracy Mean:	0.0053	0.0040	0.0066	0.0039	0.0252	0.0063	0.0200	0.0092	0.0219	0.0149	0.0021	0.0109
	Macro-Averaged Recall Score:	0.7194	0.9367	0.6541	0.9656	0.6920	0.7219	0.6524	0.9254	0.6415	0.6734	0.9320	0.7740
	Weighted Precision Score:	0.8690	0.9638	0.8502	0.9627	0.8928	0.8930	0.9477	0.9364	0.7827	0.8463	0.9503	0.8995
	Macro-Averaged Precision Score:	0.9120	0.9709	0.6421	0.8776	0.7906	0.8528	0.8814	0.9500	0.7713	0.8288	0.9189	0.8542
	Weighted F1 Score:	0.8505	0.9604	0.8691	0.9538	0.8815	0.8785	0.7807	0.9251	0.7422	0.8105	0.9495	0.8729
	Macro-Averaged F1 Score:	0.9120	0.9709	0.6421	0.8776	0.9308	0.9440	0.8953	0.9374	0.8072	0.8882	0.9189	0.8840
RF	Average Accuracy (Weighted Recall) Score:	0.8635	0.9716	0.8141	0.9420	0.9460	0.8939	0.7883	0.7133	0.7965	0.8154	0.9442	0.8626
	Standard Deviation of Accuracy Score:	0.0051	0.0031	0.0241	0.0018	0.0014	0.0080	0.0032	0.0075	0.0052	0.0083	0.0030	0.0064
	95% Confidence Interval of Accuracy Mean:	0.0032	0.0019	0.0150	0.0011	0.0009	0.0050	0.0020	0.0046	0.0032	0.0052	0.0019	0.0040
	Macro-Averaged Recall Score:	0.7500	0.9568	0.6109	0.9591	0.7621	0.7636	0.8355	0.7855	0.8604	0.6503	0.9402	0.8068
	Weighted Precision Score:	0.8732	0.9720	0.8237	0.9616	0.9536	0.8864	0.9658	0.7989	0.8533	0.8301	0.9490	0.8971
	Macro-Averaged Precision Score:	0.9168	0.9777	0.7097	0.8545	0.9220	0.8052	0.9022	0.7710	0.8106	0.8624	0.8964	0.8571
	Weighted F1 Score:	0.8545	0.9701	0.7888	0.9456	0.9310	0.8859	0.7806	0.6987	0.7870	0.7926	0.9457	0.8528
	Macro-Averaged F1 Score:	0.9168	0.9777	0.7097	0.8545	0.9220	0.8052	0.9022	0.7710	0.8106	0.8624	0.8964	0.8571
ET	Average Accuracy (Weighted Recall) Score:	0.8597	0.9677	0.8710	0.9499	0.9456	0.9178	0.7865	0.7562	0.8025	0.7959	0.9511	0.8731
	Standard Deviation of Accuracy Score:	0.0055	0.0014	0.0161	0.0015	0.0049	0.0068	0.0038	0.0093	0.0076	0.0028	0.0011	0.0055
	95% Confidence Interval of Accuracy Mean:	0.0034	0.0009	0.0100	0.0009	0.0030	0.0042	0.0023	0.0057	0.0047	0.0017	0.0007	0.0034
	Macro-Averaged Recall Score:	0.7482	0.9471	0.6631	0.9652	0.7153	0.7185	0.7526	0.8223	0.8369	0.6215	0.9438	0.7940
	Weighted Precision Score:	0.8718	0.9688	0.8794	0.9692	0.9614	0.9084	0.9987	0.8294	0.8478	0.8208	0.9530	0.9099
	Macro-Averaged Precision Score:	0.9159	0.9752	0.8347	0.8626	0.9576	0.8834	0.9688	0.8115	0.8099	0.8665	0.9171	0.8912
	Weighted F1 Score:	0.8521	0.9671	0.8484	0.9546	0.9281	0.8955	0.7607	0.7473	0.7947	0.7733	0.9512	0.8612
	Macro-Averaged F1 Score:	0.9159	0.9752	0.8347	0.8626	0.9576	0.8834	0.9688	0.8115	0.8099	0.8665	0.9171	0.8912

Table A.3: Model results when trained and tested on data from the same experiment

ML Model:	Training Data: / Testing Data:	1A Set 1 / 1A Set 2	2B Set 1 / 2B Set 2	3C Set 1 / 3C Set 2	4D Set 1 / 4D Set 2	Average:
DT	Average Accuracy (Weighted Recall) Score:	0.9923	0.9785	0.9304	0.9931	0.9736
	Standard Deviation of Accuracy Score:	0.0037	0.0032	0.0024	0.0011	0.0026
	95% Confidence Interval of Accuracy Mean:	0.0023	0.0020	0.0015	0.0007	0.0016
	Macro-Averaged Recall Score:	0.9930	0.9841	0.8667	0.9896	0.9584
	Weighted Precision Score:	0.9923	0.9785	0.9304	0.9931	0.9736
	Macro-Averaged Precision Score:	0.9736	0.9848	0.8457	0.9872	0.9478
	Weighted F1 Score:	0.9923	0.9785	0.9304	0.9931	0.9736
	Macro-Averaged F1 Score:	0.9736	0.9848	0.8457	0.9872	0.9478
SVM	Average Accuracy (Weighted Recall) Score:	0.9829	0.9674	0.9353	0.9900	0.9689
	Standard Deviation of Accuracy Score:	0.0000	0.0000	0.0000	0.0000	0.0000
	95% Confidence Interval of Accuracy Mean:	0.0000	0.0000	0.0000	0.0000	0.0000
	Macro-Averaged Recall Score:	0.9788	0.9763	0.7760	0.9926	0.9309
	Weighted Precision Score:	0.9829	0.9674	0.9353	0.9900	0.9689
	Macro-Averaged Precision Score:	0.9259	0.9569	0.9601	0.9854	0.9571
	Weighted F1 Score:	0.9829	0.9674	0.9353	0.9900	0.9689
	Macro-Averaged F1 Score:	0.9259	0.9569	0.9601	0.9854	0.9571
kNN	Average Accuracy (Weighted Recall) Score:	0.9915	0.9902	0.9496	0.9900	0.9803
	Standard Deviation of Accuracy Score:	0.0000	0.0000	0.0000	0.0000	0.0000
	95% Confidence Interval of Accuracy Mean:	0.0000	0.0000	0.0000	0.0000	0.0000
	Macro-Averaged Recall Score:	0.9958	0.9928	0.9068	0.9777	0.9683
	Weighted Precision Score:	0.9915	0.9902	0.9496	0.9900	0.9803
	Macro-Averaged Precision Score:	0.9897	0.9930	0.8841	0.9849	0.9629
	Weighted F1 Score:	0.9915	0.9902	0.9496	0.9900	0.9803
	Macro-Averaged F1 Score:	0.9897	0.9930	0.8841	0.9849	0.9629
ANN	Average Accuracy (Weighted Recall) Score:	0.9869	0.9828	0.9475	0.9855	0.9757
	Standard Deviation of Accuracy Score:	0.0145	0.0025	0.0094	0.0036	0.0075
	95% Confidence Interval of Accuracy Mean:	0.0090	0.0016	0.0058	0.0022	0.0046
	Macro-Averaged Recall Score:	0.9638	0.9853	0.8851	0.9810	0.9538
	Weighted Precision Score:	0.9869	0.9828	0.9475	0.9855	0.9757
	Macro-Averaged Precision Score:	0.9402	0.9684	0.8956	0.9757	0.9450
	Weighted F1 Score:	0.9869	0.9828	0.9475	0.9855	0.9757
	Macro-Averaged F1 Score:	0.9402	0.9684	0.8956	0.9757	0.9450
EHV	Average Accuracy (Weighted Recall) Score:	0.9946	0.9900	0.9451	0.9950	0.9812
	Standard Deviation of Accuracy Score:	0.0019	0.0040	0.0090	0.0013	0.0041
	95% Confidence Interval of Accuracy Mean:	0.0012	0.0025	0.0056	0.0008	0.0025
	Macro-Averaged Recall Score:	0.9941	0.9927	0.8900	0.9957	0.9681
	Weighted Precision Score:	0.9946	0.9900	0.9451	0.9950	0.9812
	Macro-Averaged Precision Score:	0.9764	0.9897	0.8777	0.9896	0.9584
	Weighted F1 Score:	0.9946	0.9900	0.9451	0.9950	0.9812
	Macro-Averaged F1 Score:	0.9764	0.9897	0.8777	0.9896	0.9584
ESV	Average Accuracy (Weighted Recall) Score:	0.9966	0.9904	0.9448	0.9912	0.9808
	Standard Deviation of Accuracy Score:	0.0028	0.0033	0.0057	0.0015	0.0033
	95% Confidence Interval of Accuracy Mean:	0.0017	0.0020	0.0035	0.0010	0.0021
	Macro-Averaged Recall Score:	0.9968	0.9930	0.8973	0.9822	0.9673
	Weighted Precision Score:	0.9966	0.9904	0.9448	0.9912	0.9808
	Macro-Averaged Precision Score:	0.9879	0.9931	0.8771	0.9858	0.9610
	Weighted F1 Score:	0.9966	0.9904	0.9448	0.9912	0.9808
	Macro-Averaged F1 Score:	0.9879	0.9931	0.8771	0.9858	0.9610
Stacked SVM	Average Accuracy (Weighted Recall) Score:	0.9883	0.9798	0.9300	0.9868	0.9712
	Standard Deviation of Accuracy Score:	0.0050	0.0038	0.0039	0.0076	0.0051
	95% Confidence Interval of Accuracy Mean:	0.0031	0.0024	0.0024	0.0047	0.0032
	Macro-Averaged Recall Score:	0.9855	0.9851	0.8696	0.9840	0.9560
	Weighted Precision Score:	0.9883	0.9798	0.9300	0.9868	0.9712
	Macro-Averaged Precision Score:	0.9470	0.9857	0.8459	0.9849	0.9409
	Weighted F1 Score:	0.9883	0.9798	0.9300	0.9868	0.9712
	Macro-Averaged F1 Score:	0.9470	0.9857	0.8459	0.9849	0.9409
RF	Average Accuracy (Weighted Recall) Score:	0.9957	0.9881	0.9550	0.9944	0.9833
	Standard Deviation of Accuracy Score:	0.0000	0.0016	0.0046	0.0022	0.0021
	95% Confidence Interval of Accuracy Mean:	0.0000	0.0010	0.0029	0.0014	0.0013
	Macro-Averaged Recall Score:	0.9947	0.9913	0.9137	0.9926	0.9731
	Weighted Precision Score:	0.9957	0.9881	0.9550	0.9944	0.9833
	Macro-Averaged Precision Score:	0.9778	0.9914	0.9022	0.9924	0.9659
	Weighted F1 Score:	0.9957	0.9881	0.9550	0.9944	0.9833
	Macro-Averaged F1 Score:	0.9778	0.9914	0.9022	0.9924	0.9659
ET	Average Accuracy (Weighted Recall) Score:	0.9959	0.9925	0.9701	0.9948	0.9883
	Standard Deviation of Accuracy Score:	0.0009	0.0015	0.0053	0.0007	0.0021
	95% Confidence Interval of Accuracy Mean:	0.0006	0.0009	0.0033	0.0005	0.0013
	Macro-Averaged Recall Score:	0.9950	0.9945	0.9543	0.9958	0.9849
	Weighted Precision Score:	0.9959	0.9925	0.9701	0.9948	0.9883
	Macro-Averaged Precision Score:	0.9789	0.9946	0.9316	0.9885	0.9734
	Weighted F1 Score:	0.9959	0.9925	0.9701	0.9948	0.9883
	Macro-Averaged F1 Score:	0.9789	0.9946	0.9316	0.9885	0.9734

Table A.4: Noisy training model generalizability analysis

ML Model:	Training Data: / Testing Data:	3C L0 6C L0	3C L0 - L1 6C L0	3C L0 - L2 6C L0	3C L0 - L3 6C L0	3C L0 - L4 6C L0	3C L0 & L2 6C L0	3C L0 & L3 6C L0	3C L0 & L4 6C L0
DT	Average Accuracy Score:	0.7834	0.8649	0.8649	0.7712	0.8176	0.8649	0.8283	0.8283
	Standard Deviation of Accuracy Score:	0.0976	0.0000	0.0000	0.0000	0.0391	0.0000	0.0058	0.0401
	95% Confidence Interval of Accuracy Mean Score:	0.0605	0.0000	0.0000	0.0000	0.0242	0.0000	0.0036	0.0249
	Average Error:	0.2166	0.1351	0.1351	0.2288	0.1824	0.1351	0.1717	0.1717
	Average Error Reduction %:	0.00	37.63	37.63	-5.63	15.79	37.63	20.72	20.72
	Error Reduction % Standard Deviation:	42.27	0.00	0.00	0.00	18.06	0.00	2.69	18.53
	95% Confidence Interval of Error Reduction Avg:	26.20	0.00	0.00	0.00	11.19	0.00	1.67	11.48
SVM	Average Accuracy Score:	0.8758	0.8780	0.8780	0.9063	0.8867	0.8780	0.8758	0.8758
	Standard Deviation of Accuracy Score:	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	95% Confidence Interval of Accuracy Mean Score:	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Average Error:	0.1242	0.1220	0.1220	0.0937	0.1133	0.1220	0.1242	0.1242
	Average Error Reduction %:	0.00	1.75	1.75	24.56	8.77	1.75	0.00	0.00
	Error Reduction % Standard Deviation:	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	95% Confidence Interval of Error Reduction Avg:	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
kNN	Average Accuracy Score:	0.8998	0.8976	0.8954	0.8954	0.8932	0.8976	0.8976	0.8998
	Standard Deviation of Accuracy Score:	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	95% Confidence Interval of Accuracy Mean Score:	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Average Error:	0.1002	0.1024	0.1046	0.1046	0.1068	0.1024	0.1024	0.1002
	Average Error Reduction %:	0.00	-2.17	-4.35	-4.35	-6.52	-2.17	-2.17	0.00
	Error Reduction % Standard Deviation:	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	95% Confidence Interval of Error Reduction Avg:	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ANN	Average Accuracy Score:	0.8455	0.8904	0.8656	0.8895	0.8667	0.8660	0.8586	0.8623
	Standard Deviation of Accuracy Score:	0.0334	0.0346	0.0259	0.0272	0.0261	0.0514	0.0317	0.0320
	95% Confidence Interval of Accuracy Mean Score:	0.0207	0.0214	0.0161	0.0169	0.0162	0.0319	0.0196	0.0198
	Average Error:	0.1545	0.1096	0.1344	0.1105	0.1333	0.1340	0.1414	0.1377
	Average Error Reduction %:	0.00	29.06	12.98	28.49	13.68	13.26	8.46	10.86
	Error Reduction % Standard Deviation:	27.61	22.39	16.77	17.63	16.92	33.31	20.52	20.72
	95% Confidence Interval of Error Reduction Avg:	17.11	13.88	10.40	10.93	10.49	20.64	12.72	12.84
EVH	Average Accuracy Score:	0.8641	0.9048	0.9065	0.8841	0.8856	0.9046	0.8619	0.8582
	Standard Deviation of Accuracy Score:	0.0180	0.0165	0.0177	0.0150	0.0113	0.0128	0.0207	0.0171
	95% Confidence Interval of Accuracy Mean Score:	0.0112	0.0102	0.0110	0.0093	0.0070	0.0079	0.0128	0.0106
	Average Error:	0.1359	0.0952	0.0935	0.1159	0.1144	0.0954	0.1381	0.1418
	Average Error Reduction %:	0.00	29.97	31.25	14.74	15.87	29.81	-1.60	-4.33
	Error Reduction % Standard Deviation:	14.35	12.12	13.03	11.03	8.30	9.42	15.24	12.61
	95% Confidence Interval of Error Reduction Avg:	8.89	7.51	8.07	6.84	5.15	5.84	9.45	7.81
EVS	Average Accuracy Score:	0.8658	0.8906	0.8952	0.8887	0.8826	0.8847	0.8608	0.8534
	Standard Deviation of Accuracy Score:	0.0194	0.0175	0.0250	0.0218	0.0229	0.0139	0.0102	0.0141
	95% Confidence Interval of Accuracy Mean Score:	0.0120	0.0108	0.0155	0.0135	0.0142	0.0086	0.0063	0.0087
	Average Error:	0.1342	0.1094	0.1048	0.1113	0.1174	0.1153	0.1392	0.1466
	Average Error Reduction %:	0.00	18.50	21.91	17.04	12.50	14.12	-3.74	-9.26
	Error Reduction % Standard Deviation:	18.60	13.04	18.66	16.21	17.06	10.33	7.63	10.47
	95% Confidence Interval of Error Reduction Avg:	11.53	8.08	11.56	10.05	10.58	6.40	4.73	6.49
Stacking SVM	Average Accuracy Score:	0.8918	0.8630	0.8577	0.7712	0.8828	0.8634	0.8519	0.8285
	Standard Deviation of Accuracy Score:	0.0158	0.0359	0.0164	0.0000	0.0155	0.0076	0.0088	0.0516
	95% Confidence Interval of Accuracy Mean Score:	0.0098	0.0223	0.0102	0.0000	0.0096	0.0047	0.0055	0.0320
	Average Error:	0.1082	0.1370	0.1423	0.2288	0.1172	0.1366	0.1481	0.1715
	Average Error Reduction %:	0.00	-26.69	-31.52	-111.48	-8.36	-26.28	-36.96	-58.51
	Error Reduction % Standard Deviation:	27.24	33.21	15.18	0.00	14.29	6.98	8.16	47.72
	95% Confidence Interval of Error Reduction Avg:	16.88	20.58	9.41	0.00	8.86	4.33	5.06	29.58
RF	Average Accuracy Score:	0.8136	0.9100	0.9112	0.9252	0.8671	0.9194	0.8501	0.8369
	Standard Deviation of Accuracy Score:	0.0336	0.0293	0.0317	0.0223	0.0125	0.0264	0.0308	0.0117
	95% Confidence Interval of Accuracy Mean Score:	0.0208	0.0181	0.0196	0.0139	0.0078	0.0164	0.0191	0.0072
	Average Error:	0.1864	0.0900	0.0888	0.0748	0.1329	0.0806	0.1499	0.1631
	Average Error Reduction %:	0.00	51.73	52.37	59.85	28.70	56.75	19.59	12.52
	Error Reduction % Standard Deviation:	15.86	15.70	16.98	11.99	6.71	14.18	16.51	6.26
	95% Confidence Interval of Error Reduction Avg:	9.83	9.73	10.53	7.43	4.16	8.79	10.23	3.88
ET	Average Accuracy Score:	0.8660	0.8850	0.8798	0.8792	0.8638	0.8837	0.8681	0.8461
	Standard Deviation of Accuracy Score:	0.0184	0.0225	0.0165	0.0151	0.0122	0.0213	0.0191	0.0143
	95% Confidence Interval of Accuracy Mean Score:	0.0114	0.0139	0.0103	0.0094	0.0075	0.0132	0.0118	0.0089
	Average Error:	0.1340	0.1150	0.1202	0.1208	0.1362	0.1163	0.1319	0.1539
	Average Error Reduction %:	0.00	14.15	10.33	9.84	-1.63	13.17	1.54	-14.88
	Error Reduction % Standard Deviation:	9.53	16.79	12.34	11.28	9.08	15.92	14.25	10.69
	95% Confidence Interval of Error Reduction Avg:	5.91	10.40	7.65	6.99	5.62	9.87	8.83	6.63
Average of All Models	Average Accuracy Score:	0.8634	0.8873	0.8832	0.8680	0.8725	0.8856	0.8623	0.8554
	Standard Deviation of Accuracy Score:	0.0105	0.0085	0.0052	0.0053	0.0075	0.0068	0.0066	0.0090
	95% Confidence Interval of Accuracy Mean Score:	0.0065	0.0052	0.0032	0.0033	0.0046	0.0042	0.0041	0.0056
	Average Error:	0.1366	0.1127	0.1168	0.1320	0.1275	0.1144	0.1377	0.1446
	Average Error Reduction %:	0.00	17.48	14.50	3.33	6.63	16.27	-0.85	-5.85
	Error Reduction % Standard Deviation:	7.72	6.19	3.82	3.89	5.48	4.99	4.84	6.56
	95% Confidence Interval of Error Reduction Avg:	4.79	3.84	2.37	2.41	3.40	3.09	3.00	4.07

REFERENCES

[1] E. J. Weller, H. M. Schrier and B. Weichbrodt, "What Sound Can Be Expected From a Worn Tool?," *Journal of Engineering for Industry,* vol. 91, no. 3, pp. 525-534, 1969.

[2] K. Goebel, *Management of Uncertainty in Sensor Validation, Sensor Fusion, and Diagnosis of Mechanical Systems using Soft Computing Techniques,* Berkeley: University of California, Berkeley, 1996.

[3] A. Kothuru, S. P. Nooka and R. Liu, "Application of audible sound signals for tool wear monitoring using machine learning techniques in end milling," *The International Journal of Advanced Manufacturing Technology,* vol. 95, pp. 3797-3808, 2018.

[4] B. Sick, "Online and indirect tool wear monitoring in turning with artificial neural networks: a review of more thana decade of research," *Mechanical Systems and Signal Processing,* vol. 16, no. 4, pp. 487-546, 2002.

[5] M. Kuntoglu, A. Aslan, D. Y. Pimenov, U. A. Usca, E. Salur, M. K. Gupta, T. Mikolajczyk, K. Giasin, W. Kaplonek and S. Sharma, "A Review of Indirect Tool Condition Monitoring Systems and Decision-Making Methods in Turning: Critical Analysis and Trends," *Sensors,* vol. 21, no. 1, p. 108, 2021.

[6] H. M. Ertunc and K. A. Loparo, "A decision fusion algorithm for tool wear condition monitoring in drilling," *International Journal of Machine Tools & Manufacture,* vol. 41, p. 1347–1362, 2001.

[7] L. Dan and J. Mathew, "Tool wear and failure monitoringtechniques for turning-- a review," *International Journal of Machine Tools and Manufacture,* vol. 30, no. 4, pp. 579-598, 1990.

[8] S. Kurada and C. Bradley, "A review of machine vision sensors for tool condition monitoring," *Computers in Industry,* vol. 34, no. 1, pp. 55-72, 1997.

[9] W. Li, Y. B. Guo, M. E. Barkey, C. Guo and Z. Q. Liu, "Surface Integrity and Fatigue Strength of Hard Milled Surfaces," in *ASME 2011 International Manufacturing Science and Engineering Conference*, Corvallis, Oregon, USA, 2011.

[10] B. Denkena, H. K. Tonshoff, T. Friemuth, C. Mueller, H. Zenner, F. Renner and M. Koehler, "Fatigue strength of hard turned components," in *Proceedings of the 1st lnternational Conference on Manufacturing Engineering*, Sani, Halkidiki, Greece , 2002.

[11] A. Ghasempoor, T. N. Moore and J. Jeswiet, "On-line wear estimation using neural networks," *Proceedings of the Institution of Mechanical Engineers: Journal of Engineering Manufacture, Part B,* vol. 212, no. 2, pp. 105-112, 1998.

[12] N. Ambhore, D. Kamble, S. Chinchanikar and V. Wayal, "Tool condition monitoring system: A review," in *Materials Today: Proceedings*, 2015.

[13] S. Binsaeid, S. Asfour, S. Cho and A. Onar, "Machine ensemble approach for simultaneous detection of transient and gradual abnormalities in end milling using multisensor fusion," *Journal of Materials Processing Technology,* vol. 209, p. 4728–4738, 2009.

[14] J. Yuan, L. Liu, Z. Yang and Y. Zhang, "Tool Wear Condition Monitoring by Combining Variational Mode Decomposition and Ensemble Learning," *Sensors,* vol. 20, no. 21, 2020.

[15] Y. Wu, G. S. Hong and W. S. Wong, "Prognosis of the probability of failure in tool condition monitoring application-a time series based approach," *International Journal of Advanced Manufacturing Technology,* vol. 76, p. 513–521, 2015.

[16] A. Farias, S. L. R. Almeida, S. Delijaicov, V. Seriacopi and E. C. Bordinassi, "Simple machine learning allied with data-driven methods for monitoring tool wear in machining processes," *The International Journal of Advanced Manufacturing Technology,* vol. 109, p. 2491–2501, 2020.

[17] R. L. Kegg, "One-Line Machine and Process Diagnostics," *CIRP Annals,* vol. 33, no. 2, pp. 469-473, 1984.

[18] S. S. Rangwala and D. A. Dornfield, "Learning and Optimization of Machning Operations Using Computing Abilities of Neural Networks," *IEEE Transactions on Systems, Man, and Cybernetics,* vol. 19, no. 2, pp. 299-314, 1989.

[19] C. Scheffer and P. S. Heyns, "An industrial tool wear monitoring system for interrupted turning. Mechanical Systems and Signal Processing," *Mechanical Systems and Signal Processing,* vol. 18, no. 5, pp. 1219-1242, 2004.

[20] T. Xi, I. M. Beninca, S. Kehne, M. Fey and C. Brecher, "Tool wear monitoring in roughing and finishing processes based on machine internal data," *The International Journal of Advanced Manufacturing Technology,* vol. 113, p. 3543–3554, 2021.

[21] C. Liu, G. F. Wang and Z. M. Li, "Incremental learning for online tool condition monitoring using Ellipsoid ARTMAP network model," *Applied Soft Computing,* vol. 35, pp. 186-198, 2015.

[22] A. Kothuru, S. P. Nooka and R. Liu, "Cutting Process Monitoring System Using Audible Sound Signals and Machine Learning Techniques: An Application to End Milling," in *International Manufacturing Science and Enginerering Conference*, Los Angeles, California, USA, 2017.

[23] A. Kothuru, *Application of Audible Signals in Tool Condition Monitoring using Machine Learning Techniques,* Rochester: RIT Scholar Works, 2017.

[24] R. Teti, K. Jemieliak, G. O'Donnell and D. Dornfeld, "Advanced Monitoring of Machining Operations," *CIRP Annals - Manufacturing Technology,* vol. 59, pp. 717-739, 2010.

[25] C. S. Leem, D. A. Dornfeld and S. E. Dreyfus, "A Customized Neural Network for Sensor Fusion in On-Line Monitoring of Cutting Tool Wear," *Journal of Engineering for Industry,* vol. 117, no. 2, pp. 152-159, 1995.

[26] C. Druillet, J. Karandikar, C. Nath, A.-C. Journeaux, M. E. Mansori and T. Kurfess, "Tool life predictions in milling using spindle power with the neural network technique," *Journal of Manufacturing Processes,* vol. 22, pp. 161-168, 2016.

[27] K. Jemielniak, "Commercial Tool Condition Monitoring Systems," *The International Journal of Advanced Manufacturing Technology,* vol. 15, pp. 711-721, 1999.

[28] N. Ghosh, Y. B. Ravi, A. Patra, S. Mukhopadhyay, S. Paul, A. R. Mohanty and A. B. Chattopadhyay, "Estimation of tool wear during CNC milling using neural network-based sensor fusion," *Mechanical Systems and Signal Processing,* vol. 21, no. 1, pp. 466-479, 2007.

[29] Z. Li, R. Liu and D. Wu, "Data-Driven Smart Manufacturing: Tool Wear Monitoring with Audio Signals and Machine Learning," *Journal of Manufacturing Processes,* vol. 48, pp. 66-76, 2019.

[30] D. R. Salgado and F. J. Alonso, "An approach based on current and sound signals for in-process tool wear monitoring," *International Journal of Machine Tools and Manufacture,* vol. 47, no. 14, pp. 2140-2152, 2007.

[31] R. Corne, C. Nath, M. Mansori and T. Kurfess, "Study of spindle power data with neural network for predicting real-time tool wear/breakage during inconel drilling," *Journal of Manufacturing Systems,* vol. 42, no. 2, pp. 287-295, 2017.

[32] M. Schwenzer, K. Miura and T. Bergs, "Machine Learning for Tool Wear Classification in Milling Based on Force and Current Sensors," in *IOP Conference Series: Materials Science and Engineering*, 2019.

[33] S. Yin, C. Liu, Z. Zhang, Y. Lin, D. Wang, J. Tejedor, T. F. Zheng and Y. Li, "Noisy training for deep neural networks in speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing,* vol. 2, 2015.

[34] C. M. Bishop, "Training with Noise is Equivalent to Tikhonov Regularization," *Neural Computation,* vol. 7, no. 1, p. 108–116, 1995.

[35] G. An, "The Effects of Adding Noise During Backpropagation Training on a Generalization Performance," *MIT Press,* vol. 8, no. 3, pp. 643-674, 1996.

[36] F. W. Taylor, "On the Art of Cutting Metals," *Transactions of the ASME,* vol. 28, pp. 31 - 279, 1907.

[37] B. N. Colding, "A Three-Dimensional, Tool-Life Equation— Machining Economics," *Journal of Engineering for Industry,* vol. 81, no. 3, pp. 239-249, 1959.

[38] R. Woxen, "A Theory and an Equation for the Life of Lathe Tools," *Handling,* vol. 119, 1932.

[39] W. W. Gilbert, "Economics of Machining," *American Society for Metals,* Vols. Machining-Theory and Practice, pp. 465-485, 1950.

[40] M. Wang and J. Wang, "CHMM for tool condition monitoring and remaining useful life prediction," *International Journal of Advanced Manufacturing Technology,* vol. 59, p. 463–471, 2012.

[41] T. Benkedjouh, N. Zerhouni and S. Rechak, "Tool Wear Condition Monitoring based on Continuous Wavelet Transform and Blind Source Separation," *International Journal of Advanced Manufacturing Technology,* vol. 97, pp. 3311-3323, 2018.

[42] Z. Huang, J. Zhu, J. Lei, X. Li and . F. Tian, "Tool wear predicting based on multi-domain feature fusion by deep convolutional neural network in milling operations," *Journal of Intelligent Manufacturing,* vol. 31, p. 953–966, 2020.

[43] Y. Zhou, B. Sun, W. Sun and Z. Lei, "Tool wear condition monitoring based on a two-layer angle kernel extreme learning machine using sound sensor for milling process," *Journal of Intelligent Manufacturing,* 2020.

[44] Y. Lei, Z. He, Y. Ze and X. Chen, "New clustering algorithm-based fault diagnosis using compensation distance evaluation technique," *Mechanical Systems and Signal Processing,* vol. 22, p. 419–435, 2008.

[45] M. A. Saucedo-Espinosa, H. J. Escalante and A. Berrones, "Detection of defective embedded bearings by sound analysis: a machine learning approach," *Journal of Intelligent Manufacturing,* vol. 28, p. 489–500, 2017.

[46] A. Muniyappa, "Fault diagnosis of antifriction bearings through sound signals using support vector machine," *Journal of Vibroengineering,* vol. 14, no. 4, 2012.

[47] P. K. Kankar, S. C. Sharma and S. P. Harsha, "Fault diagnosis of ball bearings using machine learning methods," *Expert Systems with Applications,* vol. 38, no. 3, pp. 1876-1886 , 2011.

[48] G. Wang, Y. Yang and Z. Li, "Force Sensor Based Tool Condition Monitoring Using a Heterogeneous Ensemble Learning Model," *Sensors,* vol. 14, no. 11, pp. 21588-21602, 2014.

[49] V. Riego, M. Castejon-Limas, L. Sanchez-Gonzalez, L. Fernandez-Robles, H. Perez, J. Diez-Gonzalez and A.-M. Guerrero-Higueras, "Strong classification system for wear identification on milling processes using computer vision and ensemble learning," *Neurocomputing,* 2020.

[50] Y. Hui, X. Mei, G. Jiang, T. Tao, C. Pei and Z. Ma, "Milling Tool Wear State Recognition by Vibration Signal Using a Stacked Generalization Ensemble Model," *Shock and Vibration,* 2019.

[51] E. Kannatey-Asibu, J. Yum and T. H. Kim, "Monitoring tool wear using classifier fusion," *Mechanical Systems and Signal Processing,* vol. 85, pp. 651-661, 2017.

[52] K. Javed, R. Gouriveau, X. Li and N. Zerhouni, "Tool wear monitoring and prognostics challenges: a comparison of connectionist methods toward an adaptive ensemble model," *Journal of Intelligent Manufacturing,* vol. 29, p. 1873–1890, 2018.

[53] M. A. Mannan, A. A. Kassim and M. Jing, "Application of Image and Sound Analysis Techniques to Monitor the Condition of Cutting Tools," *Pattern Recognition Letters,* vol. 21, no. 11, pp. 969-979, 2000.

[54] D. Wu, C. Jennings, J. Terpenny and S. Kumara, "Cloud-based machine learning for predictive analytics: Tool wear prediction in milling," in *IEEE International Conference on Big Data*, 2016.

[55] *ISO8688-2 Tool life testing in milling-- Part 2: End milling,* Geneva, Switzerland: The International Organization for Standardization (ISO), 1989.

[56] R. Serra, A. Ouahabi and W. Rmili, "Evaluation of Cutting Tool Wear during Lathe Dry Turning Process from Accelerometer Data," in *4th International Conference on Tribology in Manufacturing Processes*, Nice, France, 2010.

[57] X. Liao, G. Zhou, Z. Zhang, J. Lu and J. Ma, "Tool wear state recognition based on GWO–SVM with feature selection of genetic algorithm," *The International Journal of Advanced Manufacturing Technology,* vol. 104, p. 1051–1063, 2019.

[58] T. L. Brzezinka , J. Rao, J. M. Paiva, J. Kohlscheen, G. S. Fox-Rabinovich, S. C. Veldhuis and J. L. Endrino, "DLC and DLC-WS2 Coatings for Machining of Aluminium Alloys," *Coatings,* vol. 9, no. 3, p. 192, 2019.

[59] R. L. Vaughn, "Modern Metals Machining Technology," *Journal of Engineering for Industry,* vol. 88, no. 1, pp. 65-71, 1966.

[60] W.-H. Hsieh, M.-C. Lu and S.-J. Chiou, "Application of Backpropogation Neural Network for Spindle Vibration-Based Tool Wear Monitoring in Micro-Milling," *The International journal of Advanced Manufacturing Technology,* vol. 61, pp. 53-61, 2012.

[61] M. Safavi, M. Balazinski, H. Mehmanparast and S. A. Niknam, "Experimental Characterization of Tool Wear Morphology and Cutting Force Profile in Dry and Wet Turning of Titanium Metal Matrix Composites (Ti-MMCs)," *Metals,* vol. 10, no. 11, p. 1459, 2020.

[62] H. Takeyama, Y. Doi, T. Mitsoka and H. Sekiguchi, "Sensors of tool life for optimization of machining," in *Proceedings of the 8th International Machine Tool Design and Research Conference*, 1967.

[63] S. Jetley, "Measuring cutting tool wear on-line: some practical considerations," *Manufacturing Engineering,* pp. 55-60, 1984.

[64] K. Langhammer, "Cutting forces as parameter for determining wear on carbide lathe tools and as machinability criterion for steel," *Society of Carbide and Tool Engineers,* 1976.

[65] K. Uehara, F. Kiyosawa and H. Takeshita, "Automatic tool wear monitoring in NC turning," *CIRP Annals,* vol. 28, pp. 39-42, 1979.

[66] W. Jun, C. Fuzhi and Z. Yufong, "Research on the power monitoring method for milling cutter failure," *Journal of Tsinghua University,* vol. 2, pp. 58-63, 1994.

[67] B. Kaya, C. Oysu and H. M. Ertunc, "Force-torque based on-line tool wear estimation system for CNC milling of Inconel 718 using neural networks," *Advances in Engineering Software,* vol. 42, pp. 76-84, 2011.

[68] Y. Ayed, G. Germain, A. Ammar and B. Furet, "Tool wear analysis and improvement of cutting conditions using the high-pressure water-jet assistance when machining the Ti17 titanium alloy," *Precision Engineering,* vol. 42, pp. 294-301, 2015.

[69] J. M. Lee, D. K. Choi, J. Kim and C. N. Chu, "Real-Time Tool Breakage Monitoring for NC Milling Process," *CIRP Annals,* vol. 44, no. 1, pp. 59-62, 1995.

[70] X. Li, P. K. Venuvinod and M. K. Chen, "Feed Cutting Force Estimation from the Current Measurement with Hybrid Learning," *International Journal of Advanced Manufacturing Technology,* vol. 16, p. 859–862, 2000.

[71] K. F. Martin, J. A. Brandon, B. I. Grosvenor and A. Owen, "A comparison of in-process tool wear measurement methods in turning," *Proceedings of the 26th Machine Tool Design and Research Conference,* pp. 289-296, 1986.

[72] A. Novak and G. Ossbahr, "Reliability of the cutting force monitoring in FMS-instalations," *Proceedings of the 26th Machine Tool Design and Research Conference,* pp. 325-329, 1986.

[73] Y. S. Liao, "Development of a Monitoring Technique for Tool Change Purpose in Turning Operations," in *Proceedings of the Twenty-Sixth International Machine Tool Design and Research Conference*, 1986.

[74] R. L. Lemaster, L. Lu and S. Jackson, "The use of process monitoring techniques on a CNC wood router. Part 1. Sensor selection," *Forest Products Journal,* vol. 50, no. 7/8, pp. 31-38, 2000.

[75] M. Rizal, J. A. Ghani, M. Z. Nuawi and C. H. Che Haron, "A Review of Sensor System and Application in Milling Process for Tool Condition Monitoring," *Research Journal of Applied Sciences, Engineering and Technology,* vol. 10, pp. 2083-2097, 2014.

[76] A. P. Kulkarni, G. G. Joshi, A. Karekar and V. G. Sargade, "Investigation on cutting temperature and cutting force in turning AISI 304 austenitic stainless steel using AlTiCrN coated carbide insert," *International Journal of Machining and Machinability of Materials,* vol. 15, no. 3/4, pp. 147-156, 2014.

[77] M. A. Davies, T. Ueda, R. M'Saoubi and B. Mullany, "On the measurement of temp erature in material removal processes," *Annals of the CIRP,* vol. 56, pp. 581-604, 2007.

[78] G. Barrow, "A review of experimental and theoretical techniques for assessing cutting temperatures," *CIRP Annals,* vol. 23, pp. 203-211, 1973.

[79] X. Y. Zhang, X. Lu, S. Wang, W. Wang and W. D. Li, "A multi-sensor based online tool condition monitoring system for milling process," in *51st CIRP Conference on Manufacturing Systems*, 2018.

[80] D. A. Dornfeld, "Neural Network Sensor Fusion for Tool Condition Monitoring," *CIRP Annals,* vol. 39, no. 1, pp. 101-105, 1990.

[81] J. Wang, J. Xie, R. Zhao, L. Zhang and L. Duan, "Multisensory fusion based virtual tool wear sensing for ubiquitous manufacturing," *Robotics and Computer-Integrated Manufacturing,* vol. 45, pp. 47-58, 2016.

[82] A. Jawaid, S. Koksal and S. Sharif, "Cutting performance and wear characteristics of PVD coated and uncoated carbide tools in face milling Inconel 718 aerospace alloy," *Journal of Materials Processing Technology,* vol. 116, pp. 2-9, 2001.

[83] M. Alauddin, M. A. El Baradie and M. S. J. Hashmi, "Tool-life testing in the end milling of Inconel 718," *Journal of Materials Processing Technology,* vol. 55, pp. 321-330, 1995.

[84] A. R. C. Sharman, R. C. Dewes and D. K. Aspinwall, "Tool life when high speed ball nose end milling Inconel 718," *Journal of Materials Processing Technology ,* vol. 118, pp. 29-35, 2001.

[85] I. P. Okokpujie, O. S. Ohunakin, C. A. Bolu and K. O. Okokpujie, "Experimental data-set for prediction of tool wear during turning of Al-1061 alloy by high speed steel cutting tools," *Data in Brief,* p. 1196–1203, 2018.

[86] L. V. Colwell, "Tracking tool deterioration by computer (during actual machining)," *CIRP Annals,* vol. 23, pp. 29-30, 1974.

[87] E. Kannatey-Asibu and D. A. Dornfeld, "A study of tool wear using statistical analysis of metal-cutting acoustic emission," *Wear,* vol. 76, pp. 247-261, 1982.

[88] M. S. Lan and D. A. Dornfeld, " In-process tool fracture detection," *Journal of Engineering Materials and Technology,* vol. 106, pp. 111-118, 1984.

[89] E. N. Diei and D. A. Dornfeld, "A model of tool fracture generated acoustic emission during machining," *Journal of Industrial Engineering,* vol. 109, pp. 227-233, 1987.

[90] D. A. Tobon-Mejia, K. Medjaher and N. Zerhouni, "CNC machine tool's wear diagnostic and prognostic by using dynamic Bayesian networks," *Mechanical Systems and Signal Processing,* vol. 28, p. 167–182, 2012.

[91] E. Emel and E. Kannatey-Asibu Jr., "Tool failure monitoring in turning by pattern recognition analysis of AE signals," *Journal of Industrial Engineering,* vol. 110, pp. 137-145, 1988.

[92] I. Inasaki and S. Yonetsu, " In-process detiction of cutting tool damage by acoustic emission measurement," *Proceedings of the 22nd International Machine Tool Design and Research Conference,* pp. 261-268, 1981.

[93] A. B. Sadat and S. Raman, "Detection of tool flank wear using acoustic signature analysis," *Wear,* vol. 115, pp. 265-272, 1987.

[94] L. C. Lee, "A study of noise emission for tool failure prediction," *International Journal of Machine Tool Design and Research,* vol. 26, no. 2, pp. 205-215, 1986.

[95] M. Xiao and Y. Fu, "Tool wear monitoring with microprocessor," *Research Paper of Tianjing University, China,* pp. 39-55.

[96] P. Krishnakumar, K. Rameshkumar and K. I. Ramachdran, "Tool Wear Condition Prediction Using Vibration Signals in High Speed Machining (HSM) of Titanium (Ti-6Al-4V) Alloy," *Procedia Computer Science,* vol. 50, pp. 270-275, 2015.

[97] A. Widodo and B.-S. Yang, "Support vector machine in machine condition monitoring and fault diagnosis," *Mechanical Systems and Signal Processing,* vol. 21, no. 6, pp. 2560-2574, 2007.

[98] W. Donat, K. Choi, W. An, S. Singh and K. Pattipati, "Data Visualization, Data Reduction and Classifier Fusion for Intelligent Fault Diagnosis in Gas Turbine Engines," *Journal of Engineering for Gas Turbines and Power,* vol. 130, no. 4, 2008.

[99] S. Orhan, A. O. Er, N. Camuşcu and E. Aslan, "Tool wear evaluation by vibration analysis during end milling of AISI D3 cold work tool steel with 35 HRC hardness," *NDT & E International,* vol. 40, no. 2, pp. 121-126, 2007.

[100] I. Korkut, A. Acir and A. Boy, "Application of regression and artificial neural network analysis in modelling of tool–chip interface temperature in machining," *Expert Systems with Applications,* vol. 38, pp. 11651-11656, 2011.

[101] A. Gouarir, G. Martínez-Arellano, G. Terrazas, P. Benardos and S. Ratchev, "In-Process Tool Wear Prediction System based on Machine Learning Techniques and Force Analysis," in *8th CIRP Conference on High Performance Cutting*, 2018.

[102] M. Elangovan, S. B. Devasenapati, N. R. Sakthivel and K. L. Ramachandran, "Evaluation of expert system for condition monitoring of a single point cutting tool using principle component analysis and decision tree algorithm," *Expert Systems with Applications,* vol. 38, p. 4450–4459, 2011.

[103] S. Shurrab, R. Duwairi and A. Almshnanah, "Tool Wear Prediction in Computer Numerical Control Milling Operations via Machine Learning," in *12th International Conference on Information and Communication Systems (ICICS)*, 2021.

[104] M. Castejón-Limas, L. Sánchez-González, J. Díez-González, L. Fernández-Robles, V. Riego and H. Pérez, "Texture Descriptors for Automatic Estimation of Workpiece Quality in Milling," in *Hybrid Artificial Intelligent Systems*, 2019.

[105] B. Kilundu, P. Dehombreux and X. Chiementin, "Tool wear monitoring by machine learning techniques and singular spectrum analysis," *Mechanical Systems and Signal Processing,* vol. 25, p. 400–415, 2011.

[106] G. F. Wang, Y. W. Yang, Y. C. Zhang and Q. L. Xie, "Vibration sensor based tool condition monitoring using support vector machine and locality preserving projection," *Sensors and Actuators,* vol. 209, pp. 24-32, 2014.

[107] D. D. Kong, Y. J. Chen, N. Li and S. L. Tan, "Tool wear monitoring based on kernel principal component analysis and v-support vector regression," *International Journal for Advanced Manufacturing Technology,* vol. 89, p. 175–190, 2017.

[108] A. Jegorowa, J. Górski, J. Kurek and M. Kruk, "Use of nearest neighbors (k-NN) algorithm in tool condition identification in the case of drilling in melamine faced particleboard," *Maderas Ciencia y Tecnologia,* vol. 22, no. 2, 2020.

[109] R. Sheng and X. Zhu, "Tool Wear Assessment Approach Based on the Neighborhood Rough Set Model and Nearest Neighbor Model," *Shock and Vibration,* 2020.

[110] L. Breiman, J. Friedman, C. J. Stone and R. A. Olshen, Classification and Regression Trees, Chapman and Hall/CRC, 1984.

[111] C. Cortes and V. N. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, p. 273–297, 1995.

[112] J. C. Platt, "Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods," *Advances in Large Margin Classifiers,* 1999.

[113] B. Neef, J. Bartels and S. Thiede, "Tool Wear and Surface Quality Monitoring Using High Frequency CNC Machine Tool Current Signature," in *IEEE 16th International Conference on Industrial Informatics (INDIN)*, Porto, Portugal, 2018.

[114] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review,* vol. 65, no. 6, p. 386–408, 1958.

[115] A. Siddhpura and R. Paurobally, "A review of flank wear prediction methods for tool condition monitoring in a turning process," *International Journal of Advanced Manufacturing Technology,* vol. 65, p. 371 –39, 2013.

[116] L. Wang, M. G. Mehrabi and E. Kannatey-Asibu, "Hidden Markov Model-based Tool Wear Monitoring in Turning," *Journal of Manufacturing Science and Engineering,* vol. 124, no. 3, pp. 651-658, 2002.

[117] A. Géron, Hands-On Machine Learning With Scikit-Learn And TensorFlow: Concepts, Tools, And Techniques To Build Intelligent Systems, O'Reilly Media, 2019.

[118] C. Zhang and Y. Ma, Ensemble Machine Learning: Methods and Applications, Springer Science+Business Media, LLC, 2012.

[119] J. Kittler and F. Roli, "Multiple Classifier Systems," in *Second International Workshop, MCS 2001*, 2001.

[120] I. Ren, "An Ensemble Machine Vision System for Automated Detection of Surface Defects in Aircraft Propeller Blades," Georgia Institute of Technology, Atlanta, 2020.

[121] H. Wang, Y. Yang, H. Wang and D. Chen, "Soft-Voting Clustering Ensemble," in *MCS 2013: Multiple Classifier Systems*, Nanjing, China, 2013.

[122] S. Cho, S. Binsaeid and S. Asfour, "Design of multisensor fusion-based tool condition monitoring system in end milling," *International Journal of Advanced Manufacturing Technology,* vol. 46, p. 681–694, 2010.

[123] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher and P, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research,* vol. 12, no. 85, p. 2825−2830, 2011.

[124] R. Polikar, "Ensemble Learning," in *Ensemble Machine Learning: Methods and Applications*, New York, NY, Springer Science+Business Media, LLC, 2012, pp. 1-34.

[125] J. Vasandani, S. Bharti, D. Singh, S. Priyadarshi and M. Lanham, "Improving Model Accuracy with Probability Scoring Machine Learning Models," Krannert School of Management, Purdue University, West Lafayette, IN.

[126] C. Li, W. Zhang, G. Peng and S. Liu, "Bearing Fault Diagnosis Using Fully-Connected Winner-Take-All Autoencoder," *IEEE Access,* vol. 6, pp. 6103-6115, 2018.

[127] I. A. Jones, M. P. Van Oyen, M. S. Lavieri, C. A. Andrews and J. D. Stein, "Predicting rapid progression phases in glaucoma using a soft voting ensemble classifier exploiting Kalman filtering," *Health Care Management Science,* 2021.

[128] Q. Zhou, Z. Zhang and H. Wu, "NLP at IEST 2018: BiLSTM-Attention and LSTM-Attention via Soft Voting in Emotion Classification," in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Brussels, Belgium, 2018.

[129] M. Saqlain, B. Jargalsaikhan and J. Y. Lee, "A Voting Ensemble Classifier for Wafer Map Defect Patterns Identification in Semiconductor Manufacturing," *IEEE Transactions on Semiconductor Manufacturing,* vol. 32, no. 2, pp. 171 - 182, 2019.

[130] D. H. Wolpert, "Stacked generalization," *Neural Networks,* vol. 5, no. 2, pp. 241-259, 1992.

[131] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, p. 5–32, 2001.

[132] P. Guerts, D. Ernst and L. Wehenkel, "Extremely randomized trees," *Machine Learning,* vol. 63, pp. 3-42, 2006.

[133] M. Tang, Y. Chen, H. Wu, Q. Zhao, W. Long, V. S. Sheng and J. Yi, "Cost-Sensitive Extremely Randomized Trees Algorithm for Online Fault Detection of Wind Turbine Generators," *Frontiers in Energy Research,* 2021.

[134] U. Saeed, S. U. Jan, Y.-D. Lee and I. Koo, "Fault diagnosis based on extremely randomized trees in wireless sensor networks," *Reliability Engineering & System Safety,* vol. 205, 2021.

[135] B. Zhang, "Improved Extremely Randomized Trees Model for Fault Diagnosis of Wind Turbine," *International Journal of Science,* vol. 7, no. 12, pp. 74-87, 2020.

[136] A. Paul, M. Mozaffar, Z. Yang, W.-k. Liao, A. Choudhary, J. Cao and A. Agrawal, "A Real-Time Iterative Machine Learning Approach for Temperature Profile Prediction in Additive Manufacturing Processes," in *IEEE International Conference on Data Science and Advanced Analytics*, 2019.

[137] L. P. Heck, "Signal Processing Research in Automatic Tool Wear Monitoring," *IEEE,* 1993.

[138] K. P. Zhu, Y. S. Wong and G. S. Hong, "Noise-Robust Tool Condition Monitoring in Micro-milling with Hidden Markov Models," in *Soft Computing Applications in Industry*, Springer, 2008, pp. 23-46.

[139] Y. Li, Q. Xie, H. Huang and Q. Chen, "Research on a Tool Wear Monitoring Algorithm Based on Residual Dense Network," *Symmetry,* vol. 11, p. 809, 2019.

[140] C. M. Bishop, Neural Networks for Pattern Recognition, Clarendon Press, 1996.

[141] R. Reed, Neural Smithing: Supervised Learning in Feedforward Artificial Neural Networks, Bradford Books, 1999.

[142] I. Goodfellow, Y. Bengio and A. Courville, Deep Learning, The MIT Press, 2016.

[143] J. Sietsma and R. J. F. Dow, "Creating artificial neural networks that generalize," *Neural Networks,* vol. 4, no. 1, pp. 67-79 , 1991.

[144] R. Reed, S. Oh and R. J. Marks, "Regularization using jittered training data," in *IJCNN International Joint Conference on Neural Networks*, Baltimore, MD, USA, 1992.

[145] K. Audhkhasi, O. Osoba and B. Kosco, "Noise benefits in backpropagation and deep bidirectional pre-training," *The 2013 International Joint Conference on Neural Networks (IJCNN),* 2013.

[146] B. Kosco, K. Audhkhasi and O. Osoba, "Noise can speed backpropagation learning and deep bidirectional pretraining," *Neural Networks,* vol. 129, pp. 359-384, 2020.

[147] A. Atla, R. Tada, V. Sheng and N. Singireddy, "Sensitivity of different machine learning algorithms to noise," *Journal of Computing Sciences in Colleges,* vol. 26, no. 5, p. 96–103, 2011.

[148] T. Lasota, T. Luczak, M. Niemczyk, M. Olszewski and B. Trawinski, "Investigation of Property Valuation Models Based on Decision Tree Ensembles Built over Noised Data," in *Computational Collective Intelligence: Technologies and Applications*, Craiova, Romania, 2013.

[149] T. Lasota, Z. Telec, B. Trawiński and G. Trawinski, "Investigation of Random Subspace and Random Forest Regression Models Using Data with Injected Noise," in *Knowledge Engineering, Machine Learning and Lattice Computing with Applications*, San Sebastian, Spain, 2012.

[150] M. R. Smith and T. Martinez, "The robustness of majority voting compared to filtering misclassified instances in supervised classification tasks," *Artificial Intelligence Review,* vol. 49, p. 105–130, 2018.

[151] L. C. Lu, "The Effectiveness of Various Chatter Detection Methods under Noisy Conditions," George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA, 2020.

[152] Y. Xing, Q. Song and G. Cheng, "Predictive Power of Nearest Neighbors Algorithm under Random Perturbation," in *Proceedings of the 24th International Conference on Artifi- cial Intelligence and Statistics (AISTATS)*, San Diego, California, USA, 2021.

[153] D. Aslan and Y. Altintas, "Prediction of cutting forces in 5- axis milling using feed drive current measurements," *IEEE/ASME Transactions on Mechatronics,* vol. 23, no. 2, pp. 833-844, 2018.

[154] S. W. Kim, D. W. Lee, M. C. Kang and J. S. Kim, "Evaluation of machinability by cutting environments in high-speed milling of difficult-to-cut materials," *Journal of Materials Processing Technology,* vol. 111, no. 1-3, pp. 256-26, 2001.

[155] J. Karandikar and T. Schmitz, An Analytic Framework for Optimal Milling Parameter Selection, United States, 2018.

[156] T. Salimans and D. P. Kingma, "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks," in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, Barcelona, Spain, 2016.

[157] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication,* vol. 12, pp. 247-251, 1993.

[158] W. Li, P. Fu and W. Cao, "Study on Feature Selection and Identification Method of Tool Wear States based on SVM," *International Journal on Smart Sensing and Intelligent Systems,* vol. 6, no. 2, pp. 448-465, 2013.

[159] X. Zhang, "Deep Learning Driven Tool Wear Identification and Remaining Useful Life Prediction," Coventry University, 2020.

[160] L. F. Kozachenko and N. N. Leonenko, "Sample Estimate of the Entropy of a Random Vector," *Problems of Information Transmission,* vol. 23, no. 2, pp. 9-16, 1987.

[161] B. Ross, "Mutual information between discrete and continuous data sets," *PLOS One,* vol. 9, no. 2, 2014.

[162] M. A. Hall, "Correlation-based feature selection for machine learning," The University of Waikato, Hamilton, New Zealand, 1999.

[163] X. Zhang, S. Wang, W. Li and X. Lu, "Heterogeneous sensors-based feature optimisation and deep learning for tool wear prediction," *The International Journal of Advanced Manufacturing Technology,* vol. 114, p. 2651–2675, 2021.

[164] M. Hu, W. Ming, Q. An and M. Chen, "Tool wear monitoring in milling of titanium alloy Ti–6Al–4 V under MQL conditions based on a new tool wear categorization method," *The International Journal of Advanced Manufacturing Technology,* vol. 104, p. 4117–4128, 2019.

[165] W.-N. Cheng, C.-C. Cheng, Y.-H. Lei and P.-C. Tsai, "Feature selection for predicting tool wear of machine tools," *The International Journal of Advanced Manufacturing Technology,* vol. 111, p. 1483–1501, 2020.

[166] G. Wang, Z. Guo and Y. Yang, "Force sensor based online tool wear monitoring using distributed Gaussian ARTMAP network," *Sensors and Actuators A: Physical,* vol. 192, pp. 111-118, 2013.

[167] S. Huang, X. Li and O. P. Gan, "Tool Wear Estimation using Support Vector Machines in Ball-nose End Milling," in *Annual Conference of the Prognostics and Health Management Society*, 2010.

[168] S. Ravikumar and K. I. Ramachandran, "Tool Wear Monitoring of Multipoint Cutting Tool using Sound Signal Features Signals with Machine Learning Techniques," in *Materials Today: Proceedings 5*, 2018.

[169] F. Aghazadeh, A. Tahan and M. Thomas, "Tool condition moni- toring using spectral subtraction and convolutional neural networks in milling process," *Int J Adv Manuf Technol,* vol. 98, p. 3217–3227, 2018.

[170] H. Wu, A. Lorenson, B. Anderson, L. Witteman, H. Wu, B. Meredig and D. Morgan, "Robust FCC solute diffusion predictions from ab-initio machine learning methods," *Computational Materials Science,* vol. 134, pp. 160-165, 2017.

[171] E. C. Merkle, D. Furr and S. Rabe-Hesketh, "Bayesian comparison of latent variable models: Conditional vs marginal likelihoods," *Psychometrika,* vol. 84, p. 802–829, 2019.

[172] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta and L. Ward, "Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery," *Molecular Systems Design & Engineering,* vol. 3, pp. 819-825, 2018.

[173] H.-J. Lu, N. Zou, R. Jacobs, B. Afflerbach, X.-G. Lu and D. Morgan, "Error assessment and optimal cross-validation approaches in machine T learning applied to impurity diffusion," *Computational Materials Science,* vol. 169, p. 109075, 2019.

[174] R. R. Bouckaert and E. Frank, "Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2004.

[175] C. S. Ai, Y. J. Sun, G. W. He, X. B. Ze, W. Li and K. Mao, "The Milling Tool Wear Monitoring using the Acoustic Spectrum," *The International Journal of Advanced Manufacturing Technology,* vol. 61, pp. 457-463, 2012.

[176] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, pp. 5-32, 2001.

[177] G. E. Hinton, "Connectionist Learning Procedures," *Artificial Intelligence,* vol. 40, pp. 185-234, 1989.

[178] B. S. Prasad and M. P. Babu, "Correlation between vibration amplitude and tool wear in turning:Numerical and experimental analysis," *Engineering Science and Technology, an International Journal,* vol. 20, pp. 197-211, 2017.

[179] E. Kannatey-Asibu, J. Yum and T. H. Kim, "Monitoring tool wear using classifier fusion," *Mechanical Systems and Signal Processing,* vol. 85, pp. 651-661, 2017.

[180] D. H. Wolpert, "Stacked Generalization," *Neural Networks,* vol. 5, no. 2, pp. 241-259, 1992.

[181] S. Elanayar and Y. C. Shin, "Robust Tool Wear Estimation With Radial Basis Function Neural Networks," *Journal of Dynamic Systems, Measurement, and Control,* vol. 117, pp. 459-467, 1995.

[182] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation,* vol. 10, p. 1895–1923, 1998.