Fall 2021

Neural network supervised and reinforcement learning for neurological, diagnostic, and modeling problems

Donald Wunsch III

Department:

NEURAL NETWORK SUPERVISED AND REINFORCEMENT LEARNING FOR NEUROLOGICAL, DIAGNOSTIC, AND MODELING PROBLEMS

by

DONALD COOLIDGE WUNSCH III

A THESIS

Presented to the Graduate Faculty of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

2021

Approved by:

R. Joe Stanley, Advisor
Daniel B. Hier, Co-Advisor
Steven Corns

PUBLICATION THESIS OPTION

This thesis consists of the following three articles, formatted in the style used by the Missouri University of Science and Technology:

Paper I, found on pages 3–15, Subsumption Reduces Dataset Dimensionality Without Decreasing Performance of a Machine Learning Classifier, has been submitted to the IEEE Engineering in Medicine and Biology Conference (EMBC) in February of 2021.

Paper II, found on pages 16–27, Utilizing Reinforcement Learning to Generate an Optimal Policy for Blood PH Regulation, has been submitted to the IEEE Engineering in Medicine and Biology Conference (EMBC) in May of 2021.

Paper III, found on pages 28–52, A Comparison of Three Feature Reduction Strategies for High Dimensionality of Neurological Datasets, is intended for submission to *IEEE Transactions on Biomedical and Health Informatics*.

ABSTRACT

As the medical world becomes increasingly intertwined with the tech sphere, machine learning on medical datasets and mathematical models becomes an attractive application. This research looks at the predictive capabilities of neural networks and other machine learning algorithms, and assesses the validity of several feature selection strategies to reduce the negative effects of high dataset dimensionality. Our results indicate that several feature selection methods can maintain high validation and test accuracy on classification tasks, with neural networks performing best, for both single class and multi-class classification applications. This research also evaluates a proof-of-concept application of a deep-Q-learning network (DQN) to model the impact of altered pH on respiratory rate, based on the Henderson-Hasselbalch equation. The model behaves as expected and is a preliminary example of how reinforcement learning can be utilized for medical modelling. Its sophistication will be improved in future works.

ACKNOWLEDGMENTS

I want to acknowledge the academic, moral, and financial support provided by my father and mother, Donald C Wunsch II and Hong Wunsch. They fostered my interest in machine learning and have gone above and beyond to assist me in my academic endeavors. Without their support and encouragement over the many years, I doubt I would have found this passion for machine learning in the first place.

I also sincerely thank Dr. Joe Stanley for encouraging me professionally and academically. As both my undergraduate and graduate advisor, he has facilitated and encouraged me to pursue my academic interests in medicine and machine learning.

I profoundly thank Dr. Daniel Hier for his assistance in my research efforts. He has been a stellar role model of an physician conducting research in machine learning. His constant encouragement, and meticulous guidance were instrumental to achieving my goals. I also wish to thank him for providing some of the data used in this research.

I also thank Dr. Steven Corns for his stellar teaching in his Computational Intelligence course, and for agreeing to be a member of my committee.

Lastly, I wish to thank my friends and family for the unwavering support I have received over my duration here at S&T.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

The emergence of electronic health records has led many machine learning researchers to look for applications, both for modeling and prediction in healthcare [1], [2]. Electronic health records hold vast amounts of clinical data, that is frequently highly dimensional, and incurs high computational costs and runs the risk of including redundant and irrelevant features into proposed models. Feature reduction is an attractive solution to reduce the dimensionality of such datasets. In addition, mathematical models of chemical concentrations as they relate to certain physiological processes have been developed, such as the regulation of blood pH as a function of CO_2 and HCO_3^- levels [3]. These mathematical models include a system of equations that can be integrated with other models or known behaviors of other related physiological processes. These complex models are a good fit with reinforcement learning applications, which can model dynamic physiological processes. This thesis analyzes and proposes solutions for both problems, as well as provides a future course of action and study.

Feature selection (dimension reduction) is important to machine learning applications, especially when datasets are of high dimensionality. Feature selection can improve model accuracy, reduce over-fitting, eliminate irrelevant features, reduce computation costs, and improve model interpretability [4], [5]. Multiple datasets were constructed from a dataset consisting of 364 cases, consisting of 20 neurological diseases, using three different feature selection techniques (relief filter, PCA, and subsumption). We tested 4 different algorithms (classification trees, SVM, kNN, and a multilayer perceptron (NN)) and corresponding algorithmic variations (linear SVM, quadratic SVM,

cubic SVM, etc.) and evaluated their test and 5-fold cross validation accuracies. We wanted to determine the impact of feature selection methods on validaton and test set accuracy, to adequately assess the feasibility and usability of these dimensionality reduction methods.

The second problem we looked at was a model for the impact of blood pH on respiration rate. Overall blood pH can be calculated from an application of the Henderson-Hasselbalch equation, which is a formula for the overall pH of a chemical buffer [3]. While we understand the physiological response of the body to pH changes, we do not have a quantifiable way to measure such a response. Reinforcement learning offers the potential to generate a dynamic model of the physiological response to altered blood pH. Our research developed a rudimentary model, looking only at the response of the lungs (via respiratory rate), which serves as a proof-of-concept for additional research that would require the integration of additional mathematical models including renal responses to alterations in blood pH.

PAPER

I. SUBSUMPTION REDUCES DATASET DIMENSIONALITY WITHOUT DECREASING PERFORMANCE OF A MACHINE LEARNING CLASSIFIER

Donald Wunsch III[1] and Daniel B. Hier[1]

[1]Missouri University of Science and Technology, Department of Electrical and Computer Engineering, Rolla, MO, 65409

ABSTRACT

When features in a high dimension dataset are organized hierarchically, there is an inherent opportunity to reduce dimensionality. Since more specific concepts are subsumed by more general concepts, subsumption can be applied successively to reduce dimensionality. We tested whether subsumption could reduce the dimensionality of a disease dataset without impairing classification accuracy. We started with a dataset that had 168 neurological patients, 14 diagnoses, and 293 unique features. We applied subsumption repeatedly to create eight successively smaller datasets, ranging from 293 dimensions in the largest dataset to 11 dimensions in the smallest dataset. We tested a MLP classifier on all eight datasets. Precision, recall, accuracy, and validation declined only at the lowest dimensionality. Our preliminary results suggest that when features in a high dimension dataset are derived from a hierarchical ontology, subsumption is a viable strategy to reduce dimensionality.

1. INTRODUCTION

Electronic health records (EHR) hold huge amounts of clinical data. Some of the value of this data can be unlocked by machine learning [1], [2]. It is estimated that the EHR system of a large healthcare organization holds clinical information equivalent to 100 million years of patient data (10 million patients times 10 years) [3]. Each hospital encounter generates as much as 150,000 pieces of data. Although some hospital data is numerical (e.g. laboratory results), admission notes, progress notes, and discharge summaries are difficult to convert to a computable form. One approach to making the signs and symptoms of patients computable has been called deep phenotyping. With deep phenotyping, the signs and symptoms of patients are represented as concepts from an ontology such as the Human Phenotype Ontology (HPO) [4] – [6].

Disease classification is an important goal of machine learning healthcare applications [1]. The signs and symptoms of patients are important features utilized by machine learning classifiers to make medical diagnoses. Healthcare datasets are generally of high dimensionality with hundreds or thousands of features. For example, the Human Phenotype Ontology, used to encode the signs and symptoms of subjects with human diseases, has 19,249 unique concepts, offering “a standardized set of phenotypic terms that are organized in a hierarchical fashion. Using standardized hierarchies enables us to put our phenotypic knowledge into an organized framework that can be analyzed by computational means” [7]

Feature selection (dimension reduction) is important to machine learning applications, especially for datasets of high dimensionality. Feature selection can improve

model accuracy, reduce over-fitting, eliminate irrelevant features, reduce computation costs, and improve model interpretability [8], [9]. Approaches to reducing feature dimensionality have included filter methods, wrapper methods, ensemble methods, principal components analysis, and genetic algorithms [8]–[10].

Ontologies offer a unique additional opportunity for dimension reduction due to their inherent hierarchical structure. Most medical terminology ontologies are based on a subsumptive containment hierarchy with classes hierarchically organized from the general to the specific; also known as IS-A hierarchies. Each child class inherit properties from its parent class. The inheritance of properties from a parent is called subsumption. Subsumption supports dimension reduction. For example, the children concepts micrographia, masked face, impaired turns, decreased arms swing, reduced blink rate are subsumed under the more general concept bradykinesia (Figure 1). Similarly, the concepts fine tremor, resting tremor, action tremor, postural tremor, voice tremor, senile tremor are subsumed under the more general concept tremor. The hierarchical structure of ontologies and the ability to collapse sub-classes into more general super-classes makes an ontology well-suited for feature reduction.

2. METHODS

2.1. PROPOSED APPROACH

We proposed to test the hypothesis that the hierarchical structure of ontologies can be used to reduce the dimensionality of disease datasets without an adverse impact classification accuracy. We tested this hypothesis on a disease dataset with 168 instances

(patients), 293 unique features (signs and symptoms), 1953 total features, and 14 unique labels (diagnoses). Features were derived from a hierarchical ontology with 1242 unique concepts based on the neurological examination [11], [12]. We tested classification accuracy, precision, and recall at 8 different levels of specificity within the ontology hierarchy, reflecting a reduction in dataset dimensionality from 293 to 11 dimensions.

2.2. DIMENSIONALITY REDUCTION

We used Python to traverse the neuro-ontology [11] from each of its 1242 terminal nodes to the root node. We created 1242 ordered lists (one for each concept) of length n=8 where the last element in the list was the penultimate concept (last node prior to root) and the first element in the list was the terminal concept. If the list was less than 8 elements long, it was backfilled to 8 elements by repeating the first element (terminal node) until all lists were 8 elements in length. For example, the list for micrographia (Figure 1) was [micrographia, micrographia, micrographia, micrographia, bradykinesia, hypokinesia, movement disorder, motor finding]. Using these ordered lists as a reference, we created eight new datasets by sequentially replacing the first element in the ordered list with the second element and so on, seven times. This allowed us to perform dimension reduction sequentially with each reduction reflecting replacement of a child concept with its parent concept.

2.3. MACHINE LEARNING CLASSIFIER AND METRICS

We used MATLAB to construct a multilayer perceptron (MLP) of 3 hidden layers, each with 300 neurons.

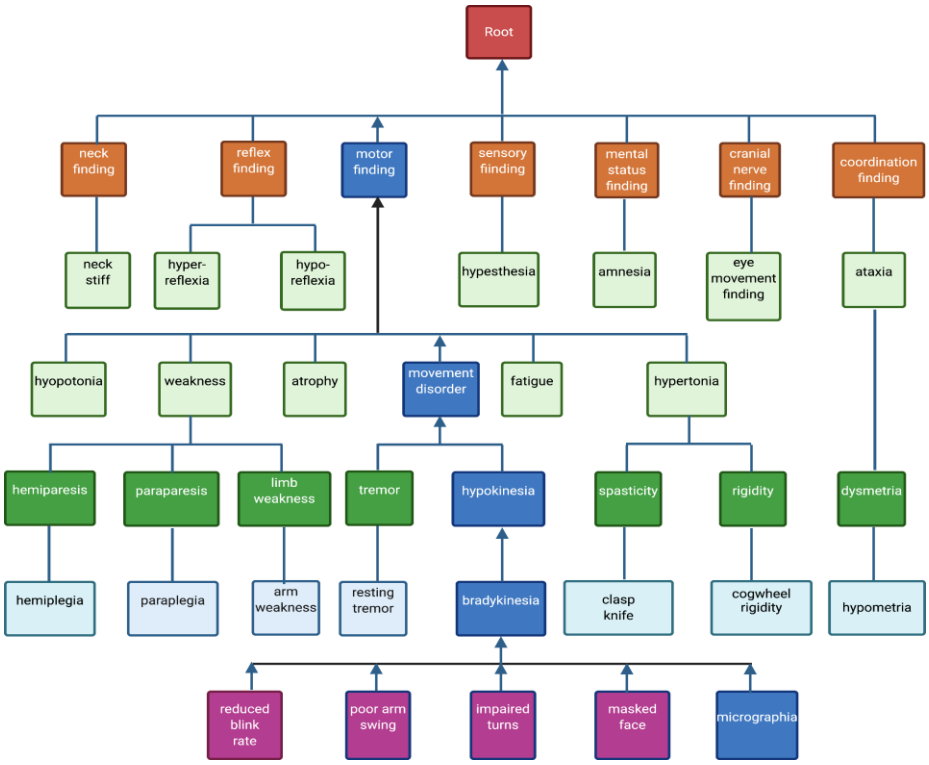

Figure 1. A small excerpt from the neuro-ontology. The neuro-ontology has 11 major branches below the root (seven shown) and 1242 terminal nodes. Concepts in the ontology become increasingly specific at lower levels going from coarsest (least specific) to most granular (most specific) at the lowest level. The concept *micrographia* (shown in dark blue) is most specific and is subsumed by *bradykinesia*, then *movement disorder*, and finally by the coarsest (least specific) concept *motor finding*.

Table 1. Diagnoses and Typical Findings.

Diagnosis	N	Finding
amyotrophic lateral sclerosis	22	weakness, fasciculations, hyperreflexia
dystonia	10	dystonia
normal pressure hydrocephalus	14	dementia, gait apraxia, incontinence
Lewy body dementia	6	dementia, hallucinations, bradykinesia
hemiballismus	4	hemiballismus
myasthenia gravis	18	weakness, diplopia, ptosis
moyopathy	18	proximal weakness
Huntington disease	17	personality change, chorea, dementia
essential tremor	7	tremor
Parkinson disease	20	tremor, bradykinesia, rigidity
multiple system atrophy	9	dysautonomia, bradykinesia, rigidity
progressive supranuclear palsy	9	gaze palsy, bradykinesia, rigidity
spinocerebellar ataxia	5	ataxia, weakness, spasticity
Wilson disease	9	tremor, ataxia, personality change

Each neuron utilized a hyperbolic tangent transfer function. Output layers used a SoftMax transfer function. The learning rate was set at 0.01 with a momentum constant of 0.1. Our dataset was split into training, testing, and validation subsets using a 70:15:15 ratio respectively. Each trial was constrained to a maximum of 1000 epochs (most trials ran for fewer than 60 epochs). Training performance was evaluated by cross-entropy, which consistently yielded higher classification accuracy than a mean-squared error performance metric [13].

Each classification was one-against-rest (OAR). The limited size of the dataset precluded meaningful classification results with some of the diagnosis classes with few members (Table I). Accuracy, precision, recall, and minimum validation loss were recorded and averaged across 10 trials at each of the eight ontology levels. Two-way ANOVA and post hoc testing were by GraphPad Prism 9.

3. RESULTS

3.1. DIMENSION REDUCTION

Using sequentially repeated subsumption based on hierarchical levels in the ontology, we reduced dimensionality from 293 dimensions to 11 dimensions (Table 2). Each case was represented by eight different vectors of successively lower dimensionality based on the hierarchy of signs and symptoms in the neuro-ontology.

3.2. CLASSIFICATION PERFORMANCE

We tested the MLP classifier on the four most common diagnoses in the dataset (amyotrophic lateral sclerosis, myopathy, myasthenia gravis, and Parkinson disease (Table I). Classification precision, accuracy, recall, and validation loss did not decline

Table 2. Dimensionality

Level	Features
1	293
2	287
3	272
4	255
5	222
6	157
7	62
8	11

until level 8 (the level that utilized the most general concepts) of the ontology (Figures 2-5). In general, the classifier performed well on all four diagnoses. Classification performance was minimally better for the diagnosis of myasthenia gravis than the other three diagnoses (Figures 2-4).

4. DISCUSSION AND CONCLUSIONS

The features of our dataset were the signs and symptoms of patients with neurological diseases. All features were categorical. Like many disease datasets, our dataset was of high dimensionality (293 different signs and symptoms) despite having

only 168 cases (Table I). High dimensionality poses difficulties for machine learning applications because of higher computational costs and the risk of including redundant or irrelevant features into the model.

The features of our dataset were derived from a subsumptive containment hierarchy [11]. In a subsumptive containment hierarchy, more specific concepts are subsumed by more general concepts. We used subsumption successively to reduce the dimensionality of our dataset from 293 dimensions to 11 dimensions. Each successive application of subsumption reduced dimensionality of the dataset and substituted a more general concepts for a more specific concept. The performance of the MLP classifier was surprisingly lossless with dimension reduction. Performance of the classifier did not drop significantly until the eighth level of the ontology which utilized the most general concepts.

At the seventh level of the ontology, dimensionality was reduced to 62 dimensions from 293 dimensions (a 79% recudction), yet overall performance of the classifier remained high (Figures 2-5).

The goal of dimension reduction methods for high dimension datasets is to find the minimal subset of features that maintains classifier accuracy and retains predicted class sizes reflective of the class sizes in the ground truth dataset upon retraining [14]–[16]. Two commonly used strategies to reduce dataset dimensionality include feature selection and feature extraction. Feature selection (filter methods, wrapper methods) emphasize algorithms that reduce the number of features into the smallest subset that accurately predict class membership [14]–[16].

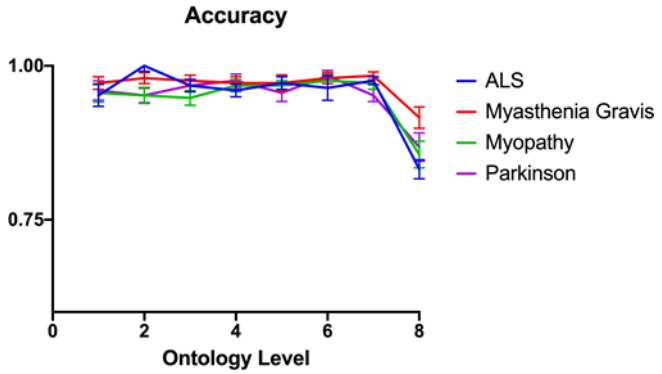

Figure 2. Accuracy (mean SEM) for classification by ontology level. Two-way ANOVA showed significant effects ($p < 0.05$) for both ontology level and diagnosis. Post hoc tests (Tukey) showed level 8 accuracy was lower than other levels and that myasthenia gravis accuracy was higher than Parkinson disease and myopathy ($p < 0.05$).

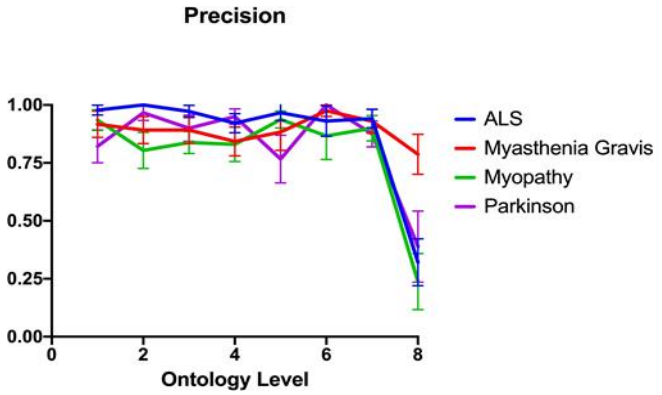

Figure 3. Precision (mean SEM) by ontology level. Two-way ANOVA showed significant effects ($p < 0.05$) for both ontology level and diagnosis. Post hoc tests (Tukey) showed level 8 precision lower than the other levels and myasthenia gravis precision higher than myopathy.

Feature extraction methods (principal components, linear discriminant analysis, etc.) emphasize methods for collapsing a large number of features into a smaller number of highly predictive features. The use of subsumption to collapse features into a smaller number of features bears more resemblance to a feature extraction strategy than a feature selection strategy. The use of knowledge embedded in a hierarchical ontology has been suggested by others as a dimension reduction strategy [17].

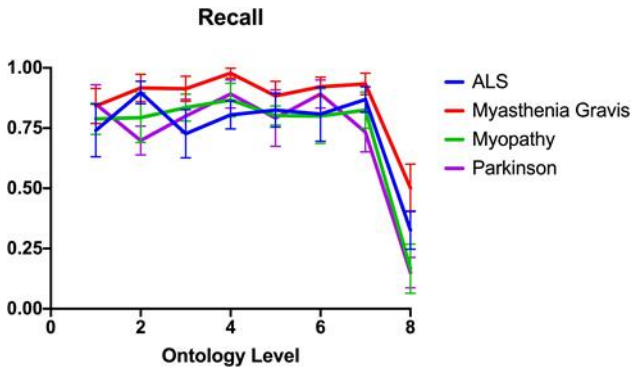

Figure 4. Recall (mean SEM) by ontology level. Two way ANOVA showed both ontology level (df= 7) and diagnosis (df=3) effects were significant ($p < 0.05$). Post hoc testing with Tukey correction showed ontology level 8 had lower recall than the other 7 levels. Recall was better for myasthenia gravis ($p<0.05$) than the other three diagnoses.

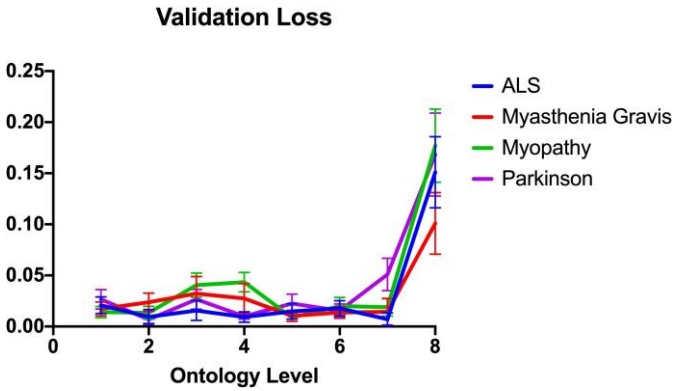

Figure 5. Validation loss (mean SEM) by ontology level. Two-way ANOVA showed ontology level was significant ($p < 0.05$). Diagnosis effect was non-significant. Post hoc comparisons with Tukey correction showed level 8 validation validation loss was higher than other levels ($P < 0.05$).

This work has important limitations. First, the dataset was small and future testing utilizing a larger dataset will be advantageous. Second, we did not test our dataset on other classifiers such as SVM, k-nearest neighbor, or logistic regression [18]. Comparison of the MLP classifier to other classifiers would be instructive. Third, due to asymmetries in the depth of the ontology, significant dimension reduction did not occur until level 5 of the ontology (Table 2). Finally, we did not compare subsumption to other

feature selections methods such as FCBF [19], mutual information [20], or Relief [21]. We plan to make these comparisons in the future. Other studies have found that when different feature reduction strategies are compared classifier performance depends on the nature of the dataset, the classifier utilized, as well as the feature reduction algorithm [18].

REFERENCES

[1] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *Journal of the American Medical Informatics Association,* vol. 25, no. 10, pp. 1419–1428, 2018.

[2] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics,* vol. 19, no. 6, pp. 1236–1246, 2018.

[3] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine,* vol. 25, no. 1, pp. 24– 29, 2019.

[4] S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Ayme´, G. Baynam, S. M. Bello, C. F. Boerkoel, K. M. Boycott et al., "The human phenotype ontology in 2017," *Nucleic acids research,* vol. 45, no. D1, pp. D865–D876, 2017.

[5] S. Köhler, N. C. Øien, O. J. Buske, T. Groza, J. O. Jacobsen, C. McNamara, N. Vasilevsky, L. C. Carmody, J. Gourdine, M. Gargano et al., "Encoding clinical data with the human phenotype ontology for computational differential diagnostics," *Current protocols in human genetics,* vol. 103, no. 1, p. e92, 2019.

[6] T. Groza, S. Köhler, D. Moldenhauer, N. Vasilevsky, G. Baynam, T. Zemojtel, L. M. Schriml, W. A. Kibbe, P. N. Schofield, T. Beck et al., "The human phenotype ontology: semantic unification of common and rare disease," *The American Journal of Human Genetics,* vol. 97, no. 1, pp. 111–124, 2015.

[7] The National Center for Biomedical Ontology, "The human phenotype ontology," https://bioportal.bioontology.org/ontologies/HP, 2021, uploaded: 2020-12-07.

[8] M. Kuhn, K. Johnson et al., *Applied predictive modeling*. Springer, 2013, vol. 26.

[9] M. Kuhn and K. Johnson, *Feature engineering and selection: A practical approach for predictive models*. CRC Press, 2019.

[10] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," in 2014 *IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 2014, pp. 1–6.

[11] D. B. Hier and S. U. Brint, "A neuro-ontology for the neurological examination," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–9, 2020.

[12] NCBO BioPortal, "Neurologic examination ontology," https://bioportal.bioontology.org/ontologies/NEO, Accessed: 2021- 01-05.

[13] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[14] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[15] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data classification: Algorithms and applications*, p. 37, 2014.

[16] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Tech. Rep., 1996.

[17] D. C. Corrales, E. Lasso, A. Ledezma, and J. C. Corrales, "Feature selection for classification tasks: Expert knowledge or traditional methods?" *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 5, pp. 2825–2835, 2018.

[18] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, "On the relationship between feature selection and classification accuracy," in *New challenges for feature selection in data mining and knowledge discovery*, 2008, pp. 90–105.

[19] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning* (ICML-03), 2003, pp. 856–863.

[20] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537–550, 1994.

[21] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, “Relief-based feature selection: Introduction and review,” *Journal of biomedical informatics*, vol. 85, pp. 189–203, 2018.

II. UTILIZING REINFORCEMENT LEARNING TO GENERATE AN OPTIMAL POLICY FOR BLOOD PH REGULATION

Donald Wunsch III

Missouri University of Science and Technology, Department of Electrical and Computer Engineering, Rolla, MO 65409

ABSTRACT

This paper assesses the feasibility of reinforcement learning for personalized modeling of the effect of respiration rate on pH regulation. It is crucial to maintaining homeostasis in the body, which is the primary function of the Bicarbonate buffer system in the bloodstream. The pH of said buffer system can be determined in terms of concentration of bicarbonate (HCO_3^-) and carbon dioxide (CO_2) by an application of the Henderson-Hasselbalch equation. We tasked a Deep-Q-Network (DQN) with maintaining physiological pH in the bloodstream in the context of respiration rate. We defined our environment in the context of four observable parameters: respiration rate, pH, concentration of CO_2 (mM), and concentration of HCO_3^- (mM). Our agent could take one of five actions: ±1 breath per minute (RR, respiration rate), ±2 RR, or maintain current RR. The trained model replicates the expected pH-regulatory behavior as a function of RR.

1. INTRODUCTION

In the absence of physiological disorders, the pH of the body ranges between 7.34 and 7.46, averaging at 7.40. This slightly alkaline pH is ideal for numerous biological functions, including the oxygenation of blood and the folding of various proteins. The importance of pH maintenance is also evidenced by numerous symptoms and disorders associated with high or low pH. This maintenance is largely achieved by chemical buffers in the body, particularly the HCO_3^- buffer. Chemical buffers are solutions comprised of a weak acid and its conjugate base, or vice versa, which have unique properties of pH change resistance; the weak acid neutralizes added base and the conjugate base neutralizes added acid [1].

Overall blood pH can be calculated from an application of the Henderson-Hasselbalch equation, which is a formula for the overall pH of a chemical buffer (Equation 1). A^- is the conjugate base and is produced because of the dissociation of the acid, HA. In the context of pH regulation via the HCO_3^- buffer, CO_2 dissolves in water to form carbonic acid, H_2CO_3, which is then converted to HCO_3^- by the actions of the enzyme carbonic anhydrase (Equation 2) [2].

$$\mathrm{pH} = \mathrm{pKa} + \log_{10}\left(\frac{[\mathrm{A}]}{[\mathrm{HA}]}\right) \tag{1}$$

$$\mathrm{CO_2} + \mathrm{H_2O} \leftrightarrow \mathrm{H_2CO_3} \xleftrightarrow{\text{Carbonic Anhydrase}} \mathrm{HCO_3^-} + \mathrm{H^+} \tag{2}$$

$$\mathrm{pH} = 6.1 + \log_{10}\left(\frac{[\mathrm{HCO_3^-}]}{[\mathrm{CO_2}]}\right) \tag{3}$$

From this reaction, we can substitute pertinent values into an application Henderson-Hasselbalch equation (Equation 3), where HCO_3^- and CO_2 concentrations are

given in mM, and the pKa value of HCO_3^- comes from the literature. This buffer system allows for pH regulation through either altering CO_2 concentration or altering HCO_3^- concentration, and we can intuitively determine the general action taken by the body for different pH disorders; decreasing CO_2 when the pH is too low, increasing CO_2 when pH is too high, increasing HCO_3^- when pH is too low, and decreasing HCO_3^- when pH is too high. These distinct actions are the responsibility of either the lungs or the kidney, respectively. Although disorders of one system can be complemented by changes to another, normal homeostasis requires these values to lie within a normal range. Disorders involving CO_2 concentration are called respiratory disorders, and disorders involving HCO_3^- are metabolic disorders [2].

While we understand the physiological background behind the body's response to bloodstream pH, we do not have an effective way to quantify just how much the body alters the rates of secretion/retention for CO_2/HCO_3^-. We seek to answer this problem by turning to machine learning, under the assumption that the most efficient policy generated using reinforcement learning may also be what the body does by extension. This paper serves as the first step towards building a reinforcement learning model that adequately represents the body's reaction in regulating blood pH. We are primarily interested in the impact of pH on the respiration rate (RR, given in breaths per minute) and the steps the body may take to maintain pH within normal limits.

2. METHODS

2.1. MODEL ASSUMPTIONS

To facilitate model construction, several assumptions are made about the body's behavior in pH regulation. The first assumption is that CO_2 exchange is continuous and always occurs. In other words, the removal of CO_2 from the bloodstream is not necessarily a discrete event (as one may intuitively believe, considering an exhale seems to be this discrete event) but is instead a continuous one. This is a fair assumption to make; if one were to hold their breath briefly, CO_2 exchange would still occur between the alveoli and the gases in their lungs.

The second assumption is a normal resting RR of 14. Most literature acknowledges the acceptable RR range from 12 to 20 breaths per minute for a resting adult, with varied changes based on age [4]. Since we also know that the standard metabolic CO_2 production of an adult on a western diet is 15000 mM/day, and that our normal resting RR must account for the removal of all this CO_2, we can intuitively define the rate of CO_2 removal at a different RR as a ratio between the new RR and the standard RR multiplied by the CO_2 decrease over a given time interval [1]. Below is the equation used to calculate the CO_2 removed in a given amount of time, T (Equation 4).

$$\mathrm{CO}_{2,\mathrm{T}}=\mathrm{CO}_{2,\mathrm{day}}\times\mathrm{K}\times\frac{\mathrm{RR}_{\mathrm{current}}}{\mathrm{RR}_{\mathrm{normal}}} \tag{4}$$

K is a unit conversion constant from day to T, $CO_{2,day}$=15000mM, and RRnormal=14 breaths per minute. In other words, we assume that CO_2 removal is linear and corresponds to our RR. This is only a first-order approximation; patients experiencing symptoms from diabetic ketoacidosis, for example, will not only experience

an increase in RR, but have deeper breaths as well [5]. While we can argue the combined impact of deeper breaths and faster rate can be approximated by assigning an even faster RR, this is an intuitive explanation and literature on the validity of such assumptions were not found. Since our model also deals with very small time intervals across each step (0.04 seconds per step), RR is treated more as a measure of the rate of CO_2 exchange rather than its typical definition of being a discrete number of breaths in a minute.

The third assumption is that pH is solely regulated by the concentration of CO_2. This is objectively not true, however, we do know that CO_2 plays a larger role in acute pH changes, as respiration is a very effective way to quickly alter arterial pH. Therefore, the assumption is appropriate for the feasibility assessment that is the goal of this paper.

2.2. AGENT AND CRITIC SPECIFICATIONS

We utilized the MATLAB Reinforcement Learning package to construct a deep-Q-network (DQN). DQNs use a deep neural network with states as input and estimated Q values as output to find a good or optimal policy directly from observations [3]. During training, the agent updates the critic properties with each step during each episode of learning and explores the action space using an epsilon-greedy exploration.

By default, the MATLAB DQN agent uses a target critic to improve the stability of the optimization through periodic updating based on the latest critic parameters. The agent was allowed to view four observations: respiration rate, pH, concentration of CO_2, and concentration of HCO_3^-. The agent was given five values to choose from as an action, all of which impact the respiration rate: ±1 RR, ±2 RR, or maintain current RR. We defined a reset function that initializes each episode with a random CO_2 and HCO_3^-value

(and thus a random pH) for each episode's start, and a step function that calculates the new pH based off the action taken, and returns said pH value alongside the HCO_3^-, CO_2, and RR. We implemented various numbers of layer sizes and determined that the best performance and behavior emerged from six layers, including and input and output layer, with two sigmoid layers alternating with two fully connected layers with thirty neurons each. The critic had a learning rate of 0.0001, and the agent was specified to update the critic every 2 steps.

2.3. REWARDS AND TRAINING SPECIFICATIONS

The reinforcement learning model was set to train up to 10000 episodes, with each episode taking up 1500 steps at most. Each step calculated a new RR and pH based off the actions of the agent and represented 0.04 seconds of modeling the CO_2 exchange. For each current observation, one of the 5 actions was selected with probability ε, which was specified as 1 at the beginning of each episode and experienced a decay rate of 0.0005. At the end of each step, ε is updated using the following formula (Equation 5).

$$\varepsilon_{new} = \varepsilon \times (1-\varepsilon_{decay}) \quad (5)$$

The maximum initial value of ε promotes exploration and helps prevent convergence on a local optimum too quickly. Rewards were given solely on the pH value at any given step of the episode; a reward of +5 was given for each step taken where the pH was between 7.41 and 7.39, close to the mean of the acceptable pH range of 7.34 and 7.46. As we approach the boundaries, however, the model receives less of a reward, going to a mere +0.1 for the ranges 7.41-7.44 and 7.36 to 7.39. In the event the pH was closer to these bounds (7.44-7.46 and 7.34-7.36) then the model was slightly punished

with a value -0.3. This addition is necessary to prevent the reinforcement learning model from accumulating rewards at a boundary without exploring the space for the better reward. Lastly, in the event the model stepped outside the normal pH bounds, the episode received a punishment of -1, and the episode prematurely ended. After all, a human with a blood pH outside of the normal range would experience severe symptoms or death. In addition to pH, CO_2 levels could also trigger a premature episode end; in the event that the respiration rate resulted in a blood CO_2 that is outside of the accepted range of 0.771-1.3566 mM, then the model would receive a punishment of -1 and end the session. These ranges are determined from literature [6].

Training could end prematurely if an average reward across 5 episodes was greater than 7000. This value corresponds to 1400 of 1500 steps per episode yielding the maximum reward. Since the CO_2 and HCO_3^- values are randomly initialized, we wanted some additional steps to allow for a model to travel from a boundary position to the maximum reward range. The theoretical maximum reward for a given episode is 7500.

3. RESULTS

3.1. TRAINING PERFORMANCE

After 1485 episodes of training, the last 5 episodes had an average reward of 7429 and ran for their maximum number of steps (1500), which triggered our stopping criteria (Figure 1.) This is a good indicator that the generated policy performs well, as each of the 5 episodes ran to completion and accrued an exceptional reward. From this policy, we

can generate simulated episodes with a smaller number of steps to take a closer look at exactly what is happening to yield this stellar result.

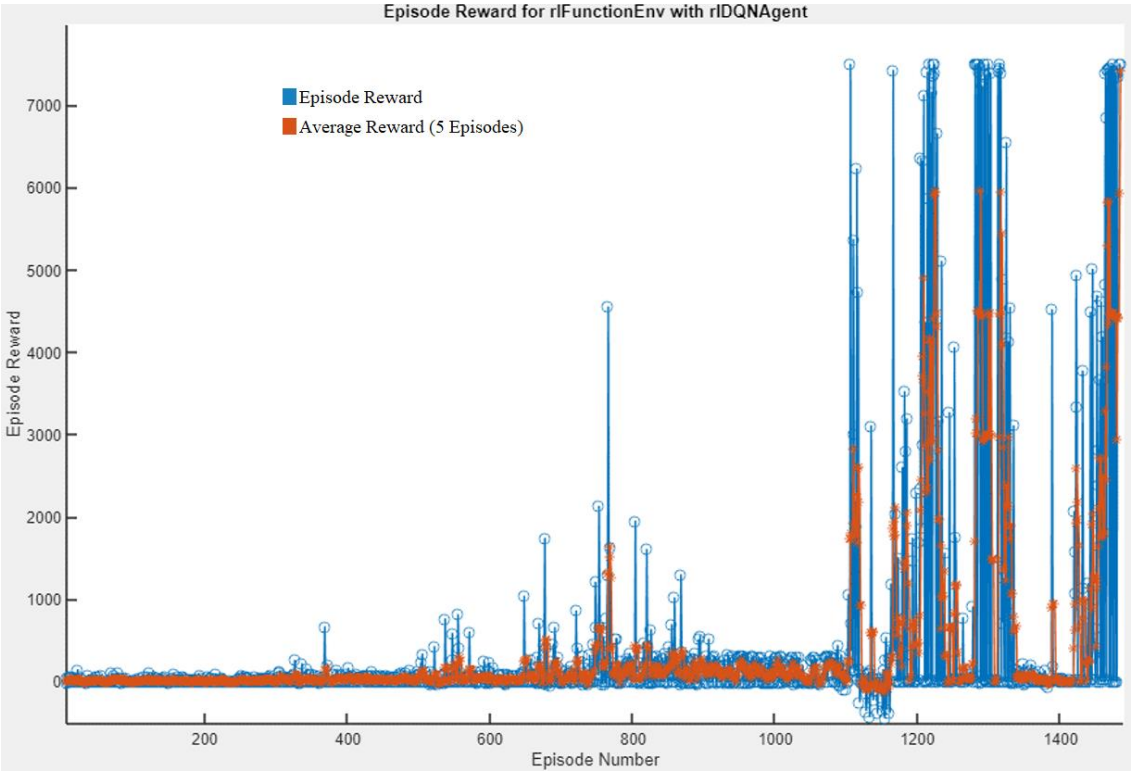

Figure 1. Total episode reward plot during training. This model generated a policy that met our stopping criterion after 1485 episodes. In blue are the rewards in each episode, and the red are average rewards over the previous 5 episodes.

3.2. POLICY VALIDATION

Running an episode for 300 steps (12 seconds) and plotting the observed environmental variables yields results as expected. As the episode progresses, steps are progressively taken to reduce the pH so that it is within our maximum reward region (7.41-7.39). In this instance, we stop at around 7.40, the exact center of this region. Once there, the policy dictates that we maintain said pH for as long as possible, and so the pH

stabilizes until the max steps is reached, which yields us the largest possible reward for the given initialization (Figure 2). Intuitively, seeing the pH decrease across our episode means we should expect a corresponding CO_2 increase (Figure 3).

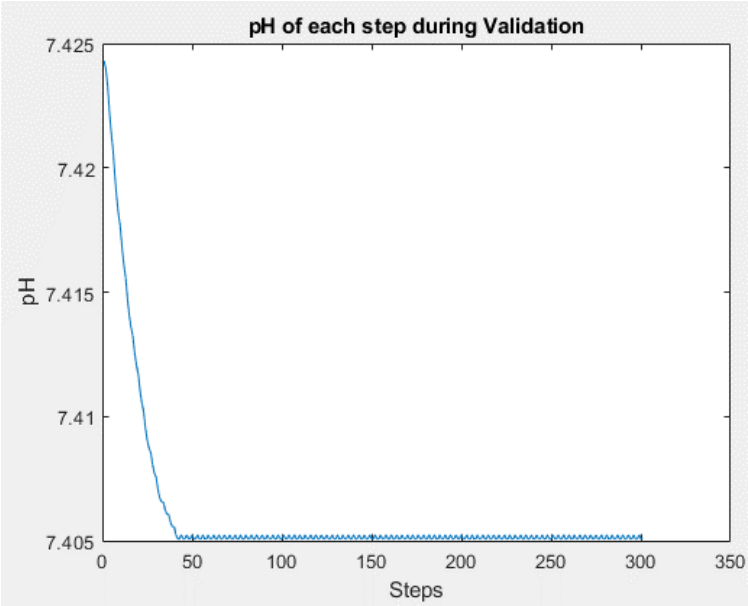

Figure 2. pH plot during validation. The policy optimized time spent in the maximum reward zone (pH = 7.41 through 7.39) and moved towards it immediately from its initial position.

Finally, the respiration rate of the validation episode behaves as expected. Since the overall CO2 of the bloodstream needs to increase, we see a rapid drop in the respiration rate from the standard value, allowing CO_2 to accumulate in the bloodstream gradually as the respiration rate is raised over time to reach the standard (Figure 4). Each step corresponds to 0.04s of CO_2 exchange, so our respiration rate is an indicator of how fast our patient is breathing over a very small period of time, not a measure of a discrete

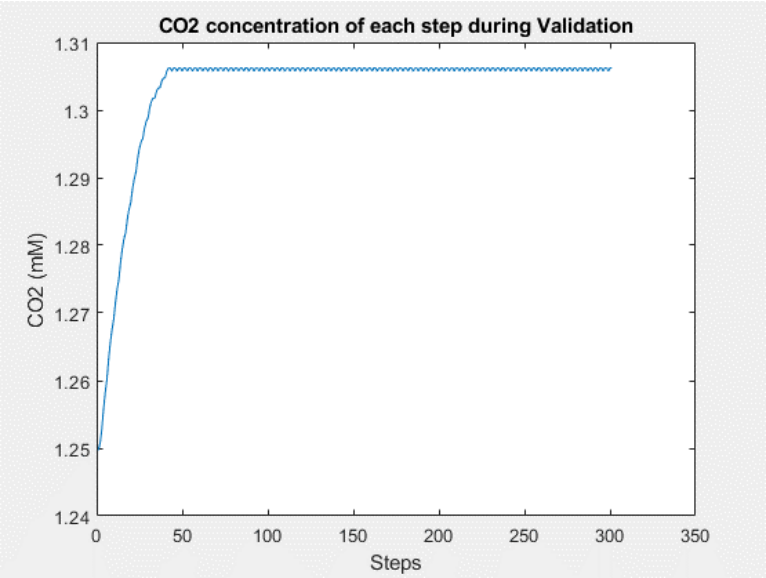

Figure 3. CO_2 plot during validation. Since the policy dictated that the pH move towards the region of highest reward, which was lower than the initial pH value, we expected the CO_2 to increase.

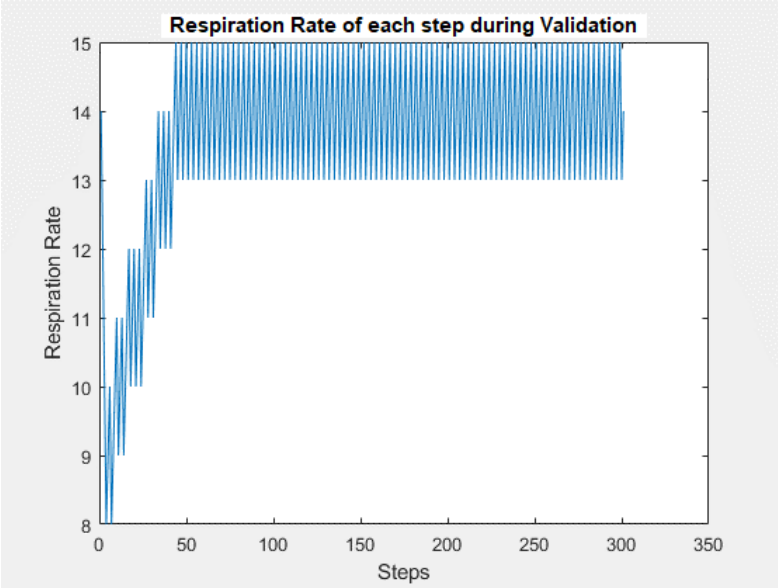

Figure 4 RR plot during validation. Since the policy dictated that the pH move towards the region of highest reward, which was lower than the initial pH value, we expected the CO2 to increase. This is accomplished by first reducing the respiration rate to below the standard so that CO2 can accumulate in the bloodstream, then gradually bring the respiration rate back up so that the CO2 is removed as fast as it is being produced.

number of breaths. In other words, RR acts as a stand in for the rate of CO_2 exchange, rather than a discrete description of the number of breaths taken in this context.

4. DISCUSSION AND CONCLUSIONS

The resulting policy generated by the reinforcement learning model behaves as expected; when pH is too high, the model will reduce the CO_2 output until we are within the range of maximum reward. This corresponds with a temporary reduction in our RR, allowing for the accumulation of CO_2, followed by a return to the standard RR, which corresponds with CO_2 leaving as fast as it is entering. Conversely, we expect to see opposing behaviors if CO_2 is too low. Thus, the model provides the ability to quantify all of these values for CO_2, HCO_3^-, RR, and pH. Further research needs to be conducted to ensure that the model is behaving as the body does.

This paper serves as the first step towards building a reinforcement learning model that adequately represents the body's reaction in regulating pH. Detailed study of this behavior in vivo must be investigated to validate and improve the reinforcement learning model. This would allow implementing additional metrics that impact overall pH, particularly HCO_3^- levels, which are primarily regulated by the liver. The regulation of pH by RR is of substantial importance for personalized medicine, particularly when therapies such as respirators have become more commonplace due to COVID-19.

ACKNOWLEDGEMENTS

I would like to thank my advisor and mentor, Daniel B. Hier, for his suggestion to pursue this research problem. I'd also like to thank my father, Donald C. Wunsch II, for his advice with the reinforcement learning aspects, and for fostering my interest in machine learning in the first place.

REFERENCES

[1] Hopkins E, Sanvictores T, Sharma S. Physiology, Acid Base Balance. [Updated 2020 Sep 14]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 Jan.

[2] Hamm LL, Nakhoul N, Hering-Smith KS. Acid-Base Homeostasis. Clin J Am Soc Nephrol. 2015;10(12):2232-2242. doi:10.2215/CJN.07400715

[3] S. Wang, H. Liu, P. H. Gomes and B. Krishnamachari, "Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks," in IEEE Transactions on Cognitive Communications and Networking, vol. 4, no. 2, pp. 257-265, June 2018, doi: 10.1109/TCCN.2018.2809722.

[4] Chourpiliadis C, Bhardwaj A. Physiology, Respiratory Rate. [Updated 2020 Sep 22]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 Jan.

[5] Lizzo JM, Goyal A, Gupta V. Adult Diabetic Ketoacidosis. [Updated 2021 Feb 6]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021 Jan.

[6] Razi E, Nasiri O, Akbari H, Razi A. Correlation of arterial blood gas measurements with venous blood gas values in mechanically ventilated patients. Tanaffos. 2012;11(4):30-35.

III. A COMPARISON OF THREE FEATURE REDUCTION STRATEGIES FOR HIGH DIMENSIONALITY NEUROLOGICAL DATASETS

Donald Wunsch III and Daniel B. Hier

Missouri University of Science and Technology, Department of Electrical and Computer Engineering, Rolla, MO, 65409

ABSTRACT

High dimensionality poses difficulties for machine learning applications because of higher computational costs and the risk of including redundant or irrelevant features into the model. Feature reduction is therefore attractive, but it is essential to maintain high accuracy, particularly for biomedical applications. We therefore investigated three different feature selection strategies; Subsumption, Relief, and Principal Component Analysis (PCA), assessing their effects on the test and validation accuracy of four representative machine learning methods. We assessed a neurological dataset containing 364 neurological patients, 20 diagnoses, and 474 unique features corresponding to signs and symptoms. We applied these feature selection strategies repeatedly to create 5 additional successively smaller datasets, ranging from our original 474 features to 11. We tested a neural network (NN), k-nearest neighbors (kNN), support vector machines (SVM), and classification trees on these various datasets to assess validation and test set accuracy. Our preliminary results suggest that Subsumption and Relief behave in similar manner with respect to validation and test set accuracy as features are reduced, and that datasets with high dimensionality can be substantially simplified (from 474 to 76) while still maintaining high accuracy. PCA, on the other hand, actuallly requires feature

reduction to perform well, quickly degrading with increasing features. Validation accuracy is high for Relief and Subsumption, but PCA has lowered average validation accuracy, indicating that there may be losses to generalizability when using this strategy.

1. INTRODUCTION AND PREVIOUS WORK

Electronic health records (EHR) hold huge amounts of clinical data. Some of the value of this data can be unlocked by machine learning [1], [2]. It is estimated that the EHR system of a large healthcare organization holds clinical information equivalent to 100 million years of patient data (10 million patients times 10 years) [3]. Each hospital encounter generates as much as 150,000 pieces of data. Although some hospital data is numerical (e.g. laboratory results), admission notes, progress notes, and discharge summaries are difficult to convert to a computable form. One approach to making the *signs and symptoms* of patients computable has been called *deep phenotyping*. With deep phenotyping, the signs and symptoms of patients are represented as concepts from an ontology such as the Human Phenotype Ontology (HPO) [4]–[6]. One of the goals of deep phenotyping is to identify disease sub phenotypes which identify the characteristics of specific subsets of a disease phenotype.

Disease classification is an important goal of machine learning healthcare applications [1]. The signs and symptoms of patients are important features utilized by machine learning classifiers to make medical diagnoses. Healthcare datasets are generally of high dimensionality with hundreds or thousands of features (Figure 1). For example, the Human Phenotype Ontology, used to encode the signs and symptoms of subjects with

human diseases, has 19,249 unique concepts, offering “a standardized set of phenotypic terms that are organized in a hierarchical fashion. Using standardized hierarchies enables us to put our phenotypic knowledge into an organized framework that can be analyzed by computational means” [7].

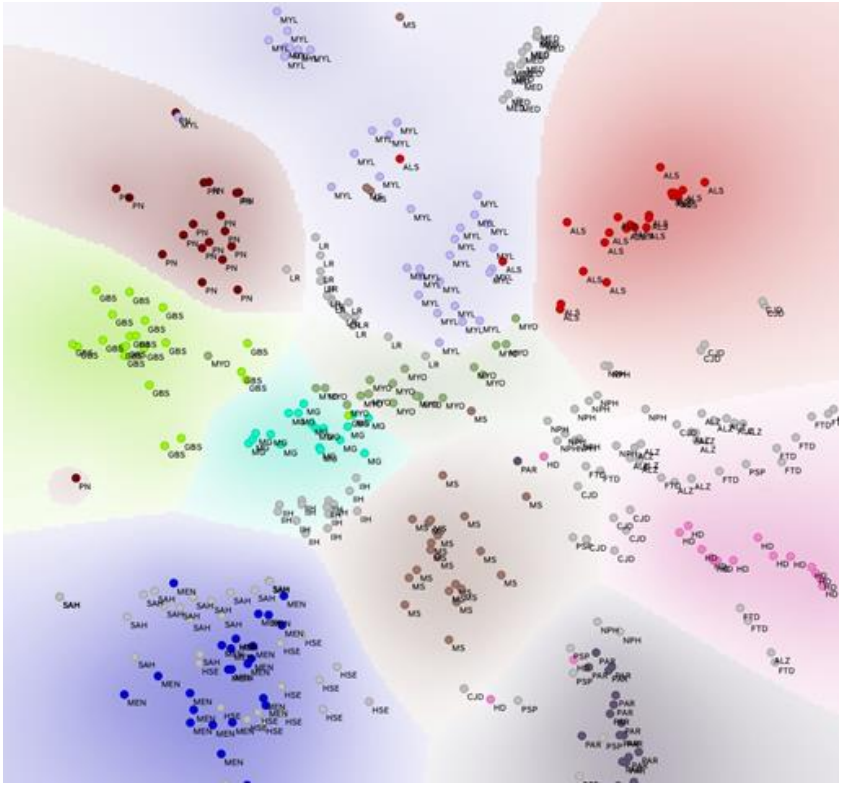

Figure 1. A t-SNE map (method by [11]) to illustrate the complexity of the diagnosis classification task. The t-SNE is based on the 20 diagnoses with each diagnosis as a different color. All 364 cases were mapped based on the full 476 feature set. If the t-SNE is viewed as a clock face, note that at 2-4 pm there is overlap between CJD, HD, ALZ, and FTD (all dementing diseases), at 5 pm there is overlap between PAR, NPH, and PSP (all hypokinetic diseases), at 7 pm there is overlap between HSE, SAH, and MEN (all meningitic diseases), at 12 noon to 1 pm there is overlap between MYL and ALS (myelopathic diseases) and in the center there is overlap between MG and MYO (pure motor diseases). For key to abbreviations and typical findings, see Table I.

Feature selection (dimension reduction) is important to machine learning applications, especially for datasets of high dimensionality. Feature selection can improve model accuracy, reduce overfitting, eliminate irrelevant features, reduce computation costs, and improve model interpretability [8], [9]. Approaches to reducing feature dimensionality have included filter methods, wrapper methods, ensemble methods, principal components analysis, and genetic algorithms [8]–[10].

Ontologies offer a unique additional opportunity for dimension reduction due to their inherent hierarchical structure. Most medical terminology ontologies are based on a *subsumptive containment hierarchy* with classes hierarchically organized from the general to the specific; also known as *IS-A hierarchies*. Each child class inherit properties from its parent class. The inheritance of properties from a parent is called *subsumption*. Subsumption supports dimension reduction. For example, the children concepts micrographia, masked face, impaired turns, decreased arms swing, reduced blink rate are subsumed under the more general concept *bradykinesia* (Figure 2). Similarly, the concepts fine tremor, resting tremor, action tremor, postural tremor, voice tremor, senile tremor are subsumed under the more general concept *tremor*. The hierarchical structure of ontologies and the ability to collapse sub-classes into more general super-classes makes an ontology well-suited for feature reduction. We use the term *subsumption* to describe this feature reduction strategy.

In this study we have compared the ability of three feature reduction strategies (*feature filtering*, *principal components*, and *subsumption*) to reduce the dimensionality of a high dimension medical dataset. Filter methods use a metric (often a distance metric between cases) to identify the best features that discriminate between cases of different

classes. Principal components analysis creates new features from a linear weighted combination of existing features that reduces dataset dimensionality without losing predictive information. Subsumption uses the hierarchical structure of an ontology to collapse more narrowly defined features into more broadly defined features.

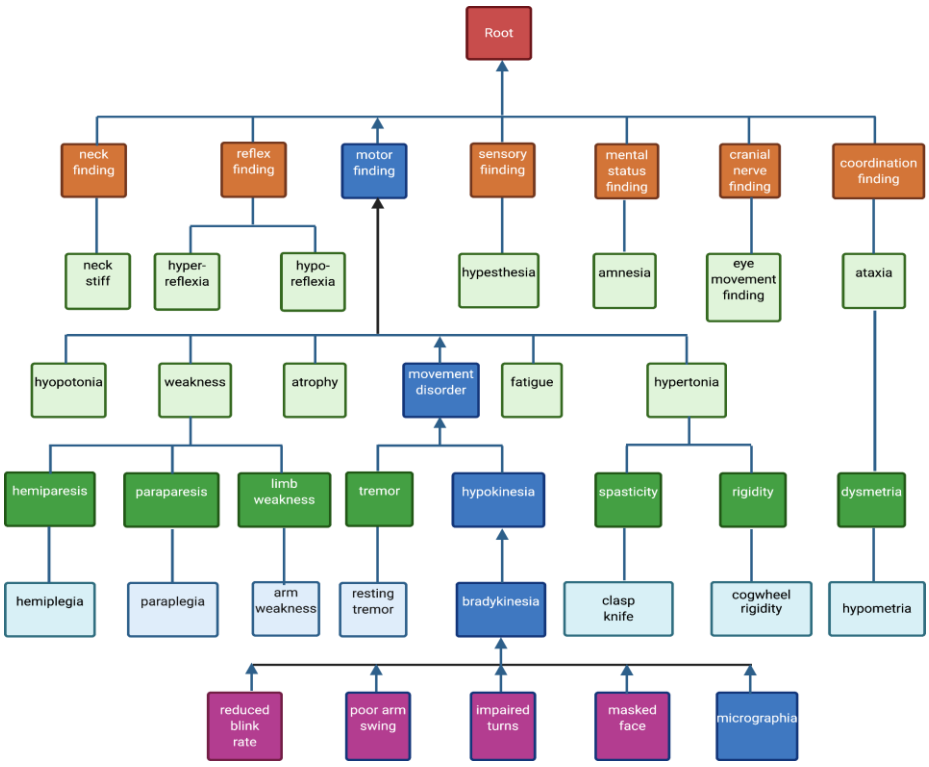

Figure 2. A small excerpt from the neuro-ontology. The neuro-ontology has 11 major branches below the root (seven shown) and 1242 terminal nodes. Concepts in the ontology become increasingly specific at lower levels going from coarsest (least specific) to most granular (most specific) at the lowest level. The concept *micrographia* (shown in dark blue) is most specific and is subsumed by *bradykinesia*, then *movement disorder*, and finally by the coarsest (least specific) concept *motor finding*. Each color represents a different level in the concept hierarchy.

The three feature reduction strategies were tested on a multi-class classification task involving 20 neurological diseases. After feature reduction, we tested classification

accuracy with four different machine learning classifiers (neural network (NN), support vector machine (SVM), k-nearest neighbor (kNN), and classification trees).

2. METHODS

2.1. OVERVIEW

We proposed to study the effects of feature selection and dimensional reduction strategies on classification accuracy across various machine learning algorithms. We tested this hypothesis on several disease datasets of varying features (ranging from 11 to 474 signs and symptoms) with 364 instances (patients) and over 20 diseases (Table 1), across 4 different algorithms (classification trees, SVM, kNN, and a multilayer perceptron (NN)) and corresponding algorithmic variations (linear SVM, quadratic SVM, cubic SVM, etc.). Multiple datasets were constructed using three different feature selection techniques (relief filter, PCA, and subsumption). Using the feature reduction strategies, new datasets of reduced dimensionality with a reduced number of features, ranging from 11 features to 464, were created. All classifiers had their test accuracy and 5-fold cross validation accuracy assessed. The mean ±s.d. of 10 trials per classifier across all conditions was calculated.

2.2. DATASET

The test dataset consisted of 364 cases, each case being a patient with one of 20 different neurological diseases (Table 1). All cases were derived from 11 standard textbooks of neurology [12]–[22]. For each entry into the dataset, the disease diagnosis

was entered as the machine learning label. Symptoms (what the patient complains of) and signs (examination findings by the physician) were abstracted from the case histories and then mapped to one of the 1404 concepts in the Neurological Examination Ontology by previously described methods [23], [24]. To capture all the signs and symptoms of the 364 cases in the dataset, 475 unique concepts were used. Each case was represented as a 476-dimension vector. The first element of the vector was the *label* (disease diagnosis) followed by 475 *features* (signs and symptoms).

Table 1. Abbreviations, Diagnoses, Counts, and Typical Findings.

Abbreviation	Diagnosis	N	Finding
ALS	amyotrophic lateral sclerosis	23	weakness, hyperreflexia, fasciculations
ALZ	Alzheimer's disease	17	dementia, memory loss
CJD	Creutzfeldt Jacob disease	12	dementia, myoclonus, ataxia
FTD	fronto-temporal dementia	13	dementia, aphasia, personality change
GBS	Guillain Barre syndrome	22	ascending weakness and numbness, hyporeflexia
HD	Huntington disease	17	personality change, dementia, chorea
HSE	herpes simplex encephalitis	16	confusion, fever, aphasia, stiff neck
IIH	idiopathic intracranial hypertension	14	headache, blurred vision, papilledema
LR	lumbar radiculopathy	16	foot weakness, sensory loss in leg, pain
MED	median nerve neuropathy	16	sensory loss in hand, pain
MEN	meningitis	24	stiff neck, fever, confusion
MG	myasthenia gravis	18	diplopia, fatiguable weakness, eyelid ptosis
MS	multiple sclerosis	24	ataxia, weakness, spasticity, optic neuritis
MYL	myelopathy	35	sensory level, Babinksi signs, weakness
MYO	myopathy	18	proximal muscle weakness
NPH	normal pressure hydrocephalus	14	urinary incontinence, dementia, gait difficulty
PAR	Parkinson disease	20	bradykinesia, rigidity, resting tremor
PN	polyneuropathy	19	weakness, sensory loss, hyporeflexia
PSP	progressive supranuclear palsy	9	bradykinesia, impaired eye movements, rigidity
SAH	subarachnoid hemorrhage	17	headache, stiff neck, vomiting
	TOTAL:	364	

All features were binarized as 0 = absent or 1 = present. The test dataset was a 364 (cases) x 476 (label + features) matrix in which all values were binary except the case labels. Cases averaged 11.2 ± 3.5 features.

2.3. DIMENSIONALITY REDUCTION

2.3.1. Dimensionality Reduction by Subsumption. The features in our dataset are concepts from the neuro-ontology [23]. The neuro-ontology is a hierarchical subsumptive ontology which supports IS-A relationships. For example, in the neuro-ontology bradykinesia IS-A hypokinesia IS-A movement disorder IS-A motor finding. Because bradykinesia is the child concept of hypokinesia, we can say that bradykinesia is subsumed by hypokinesia. Just as hypokinesia is subsumed by movement disorder (Figure 2). In a subsumptive ontology like the neuro-ontology, we can use subsumption repetitively to reduce features by consolidating all the children concepts with the parent concepts. Since the neuro-ontology is at most eight levels deep, we had a potential of 8 steps of subsumption to successively reduce dimensionality. However, some branches of the neuro-ontology were only 3 or 4 levels deep. We used Python to traverse the neuro-ontology [23] from each of its 1404 terminal nodes to the root node (Figure 1). We created 1404 ordered lists (one for each concept) of length n=8 where the last element in the list was the penultimate concept (last node prior to root) and the first element in the list was the terminal concept. If the list was less than 8 elements long, it was backfilled to 8 elements by repeating the first element (terminal node) until all lists were 8 elements in length. For example, the list for micrographia (Figure 2) was [micrographia, micrographia, micrographia, micrographia, bradykinesia, hypokinesia, movement

disorder, motor finding]. Using these ordered lists as a reference, we created eight new datasets by sequentially replacing the first element in the ordered list with the second element and so on, seven times. Two of the new datasets provided minimal feature reduction and were eliminated from the analysis. The remaining six datasets had 11, 76, 245, 360, 424, and 464 features compared to the initial dataset with 474 features.

2.3.2. Dimension Reduction by Principal Components. The term principal components was introduced by Hotelling in the 1933 [25]. It is a popular multivariate statistical technique to reduce dataset dimensionality by creating new variables that are a linear combination of existing variables. The goal of principal component analysis (PCA) is to reduce dataset dimensionality, retain as much information as possible, and to reduce noise and information redundancy [26], [27]. With PCA, the original variables are replaced with a smaller number of variables that are called factor scores (weighted linear combinations of the original variables). We used the factor analysis module of SPSS 27.0 (IBM Corporation) with extraction by principal components analysis and rotation by Varimax with Kaiser normalization to create new datasets with 11, 76, 245, 360, 424, and 464 features to parallel the dimensionality of the subsumption datasets

2.3.3. Dimension Reduction by Relief. The Relief algorithm for feature selection was originally described by Kira and Kendell [28] and later modified as ReliefF by Kononenko et al [29]. ReliefF is a filter-based feature reduction strategy that evaluates each feature independently of other features (it does not look for the best combination of features or consider redundancy between features). The algorithm is based on finding index cases in the dataset and then examining matching nearest neighbors (hits) and non-matching neighbors (misses). It then uses a difference function to looks for which

features best distinguish the hits from the misses. We used the ReliefF filter as implemented in Orange data mining [11]. To parallel the feature reductions obtained by subsumption, we used the *Ranking* widget in Orange to create 6 subsets of the original dataset with 11, 76, 245, 360, 424, and 464 features.

2.4. MACHINE LEARNING CLASSIFIERS

2.4.1. K-Nearest Neighbors. We used MATLAB to construct fine, medium, and coarse kNN classifiers corresponding to k=1, 10, and 100 nearest neighbors, respectively. We utilized the default Euclidean distance metric and standardized non-categorical predictor data. We used an 80:20 split for model training and testing, model validation was performed using 5-fold cross validation. We also ran a cosine kNN classifier corresponding to k=10 using a cosine distance metric, with the same dataset splits.

2.4.2. Support Vector Machines. We used MATLAB to construct linear, quadratic, and cubic support vector machine (SVM) classifiers. As the names imply, each SVM constructed a hyperplane boundary of order 1, 2, and 3, respectively. SVM architectures have advantages in high dimensional cases [30]. By default, the classifier uses a one-vs-one multiclass classification strategy and standardizes predictor data. We used an 80:20 split for model training and testing, model validation was performed using 5-fold cross validation.

2.4.3. Classification Trees. We used MATLAB to construct fine, medium, and coarse classification trees corresponding to various thresholds for the maximum number of splits: 100, 20, and 4, respectively. The default splitting criterion is the Gini's Diversity Index, which is standard for most decision trees [31]. We used an 80:20 split

for model training and testing, model validation was performed using 5-fold cross validation.

2.4.4. Neural Networks. We used MATLAB to construct a multilayer perceptron (MLP) of 3 hidden layers, each with 500 neurons. Each neuron utilized a hyperbolic tangent transfer function. Output layers used a SoftMax transfer function. The learning rate was set at 0.01 with a momentum constant of 0.1. Our dataset was split into training, testing, and validation subsets using a 70:15:15 ratio respectively. Each trial was constrained to a maximum of 300 epochs as a precautionary measure (most trials ran for fewer than 60 epochs). Training ceased after 6 successive increases in validation error. Training performance was evaluated by cross-entropy, which consistently yielded higher classification accuracy than a mean-squared error performance metric [32].

2.5. STATISTICAL TESTING

To test differences in group means, we used one-way ANOVA with a significance level of $p < 0.05$ (SPSS 27, IBM). Post hoc means comparisons were by the Bonferroni method.

3. RESULTS

3.1. DIMENSION REDUCTION

Using sequentially repeated subsumption based on hierarchical levels in the neuro-ontology, we created reduced dimensionality subsets from the original dataset (474 dimensions) that had 11, 76, 245, 360, 424, and 464 dimensions. Comparable datasets of

11, 76, 245, 360, 424, and 464 dimensions were created by the ReliefF filter method and the PCA method.

3.2. CLASSIFIER PERFORMANCE

We tested each of the classifiers on a multi-class classification task that involved assigning each of the 364 cases to one of 20 classes (diagnoses) based on the available features. For each classifier, the classification task was repeated on all 6 data subsets with dimensionality that ranged from 11 to 464 features.

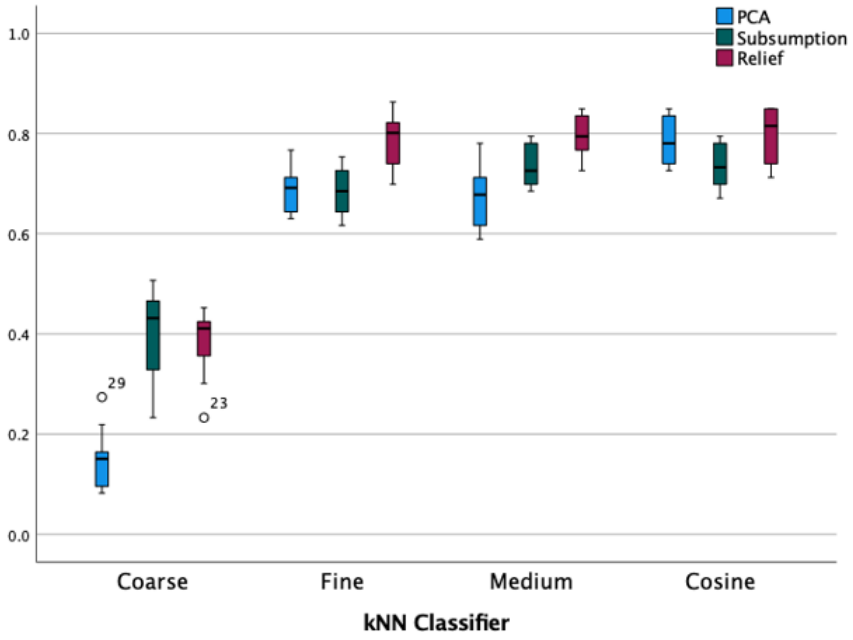

Figure 3. Comparative accuracy of four different kNN classifiers utilizing 76 features. For all three dimension reduction strategies, the four kNN classifiers performed similarly except for the coarse kNN which performed significantly worse than the other three. (One-way ANOVA with post hoc Bonferroni text, $p < 0.05$. For additional analyses, we selected the cosine kNN classifier.)

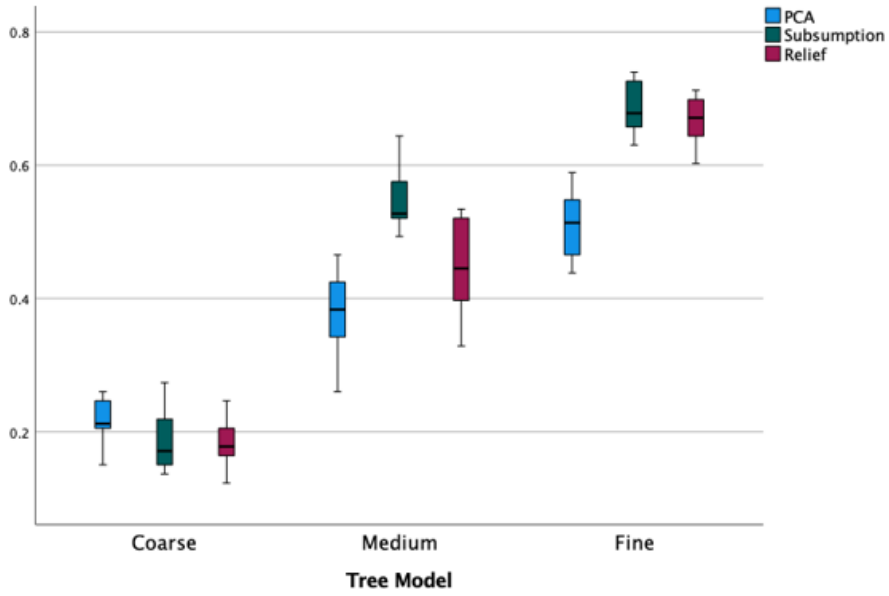

Figure 4. Comparative accuracy of three different Tree classifiers utilizing 76 features. For all three dimension reduction strategies, Fine outperformed Medium, Medium outperformed Coarse. (One-way ANOVA with post hoc Bonferroni test, p<.05. For additional analyses we selected the Fine Tree classifier.)

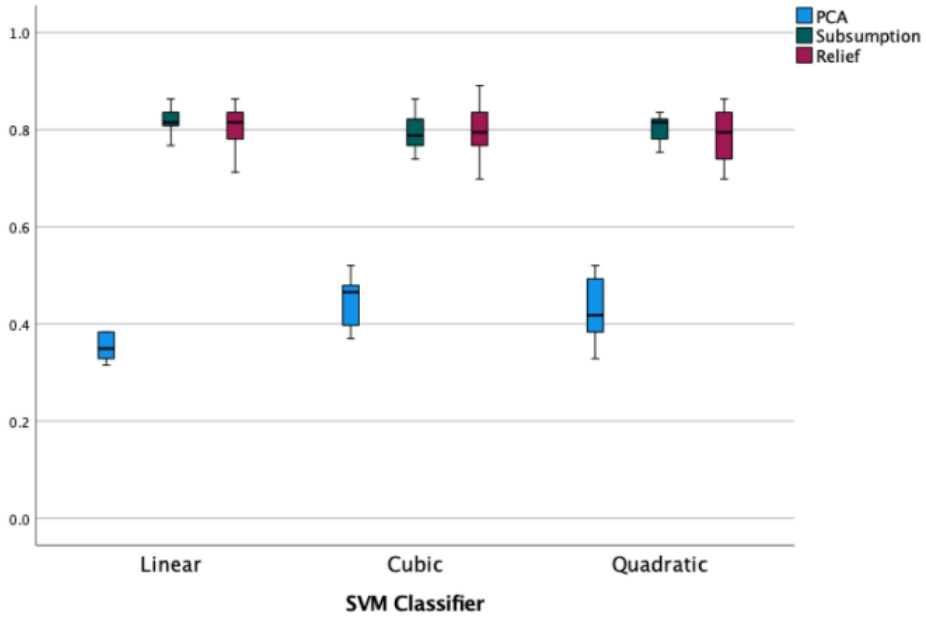

Figure 5. Comparative accuracy of three different SVM classifiers utilizing 76 features. All three SVM classifiers performed similarly, although performance was lower with the PCA feature reduction strategy (One-way ANOVA, p <0.05). For additional analyses we selected the Linear SVM classifier.)

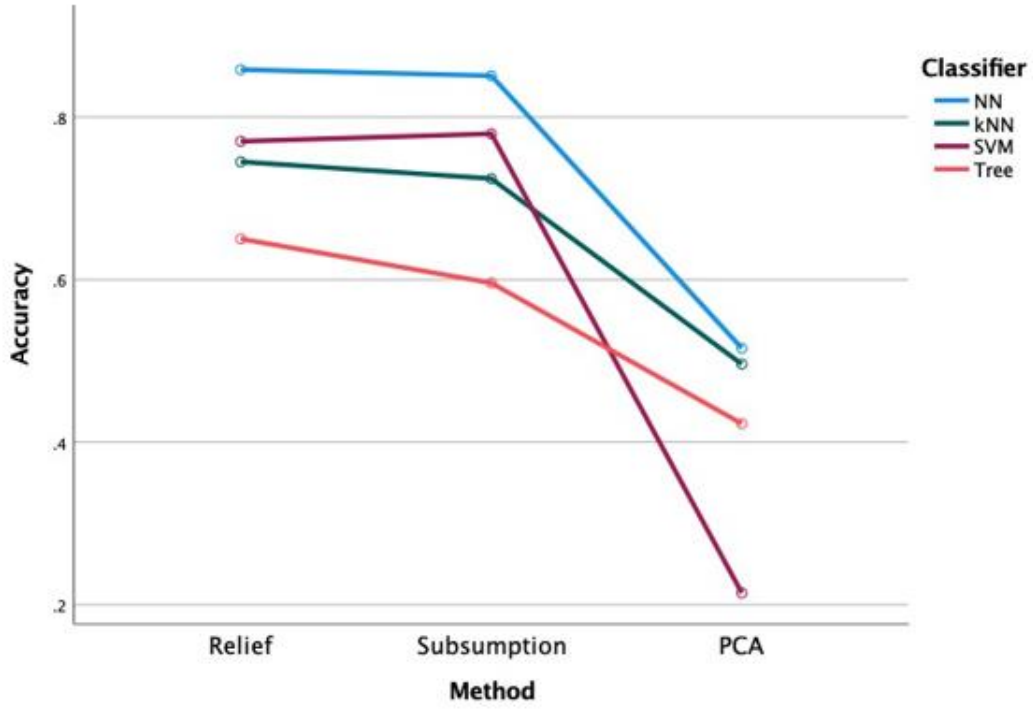

Figure 6. The average across all feature levels show that the NN classifier performed best for all three dimension reduction strategies. (One-way ANOVA, post hoc Bonferroni test, $p<0.05$). The low average accuracy for PCA for all classifiers reflects pooling of high accuracy at low dataset dimensionality with low accuracy at high dimensionality).

We first evaluated the performance of variations of the kNN, Tree, and SVM classifiers. All variations of the kNN classifier performed similarly (across all dimension reduction strategies) except for the Coarse kNN classifer which performed significantly worse than the others (Figure 3). This is likely due to the much larger value of k for the coarse classifier. We selected the Cosine kNN classifier for subsequent analyses, as it performed best. For the Tree classifiers, Fine performed better than Medium and Medium performed better than Coarse (Figure 4). The Fine Tree classifier was chosen for subsequent analyses. The linear, cubic, and quadratic versions of the SVM classifier performed similarly (Figure 5). We selected the Linear SVM classifier for subsequent analyses.

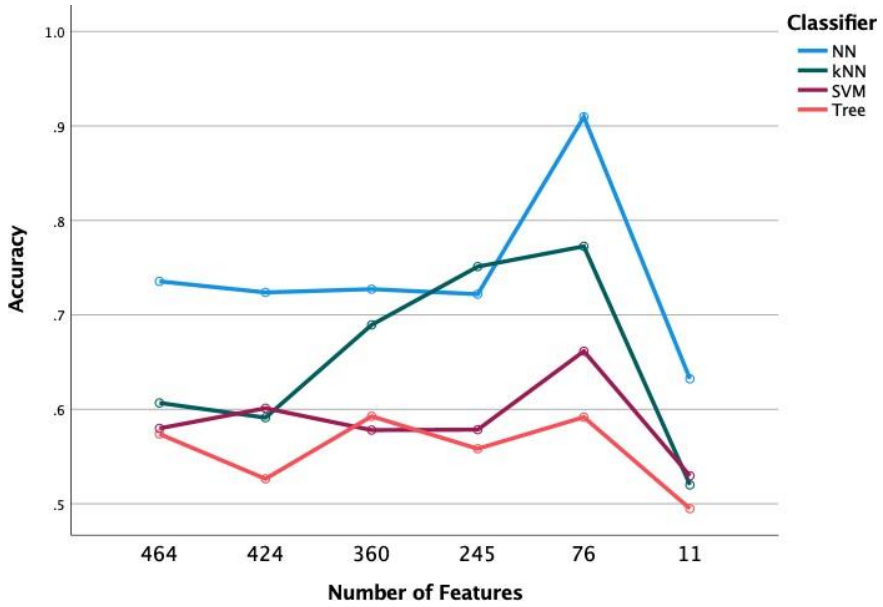

Figure 7. The NN classifier outperforms the other classifiers at all levels of dataset dimensionality, performing best near 76 features. Results are pooled across all three dimension reduction strategies. Note that subsequent figures show that dimension reduction through PCA has an opposite effect compared to relief and subsumption, lowering the depicted averages.

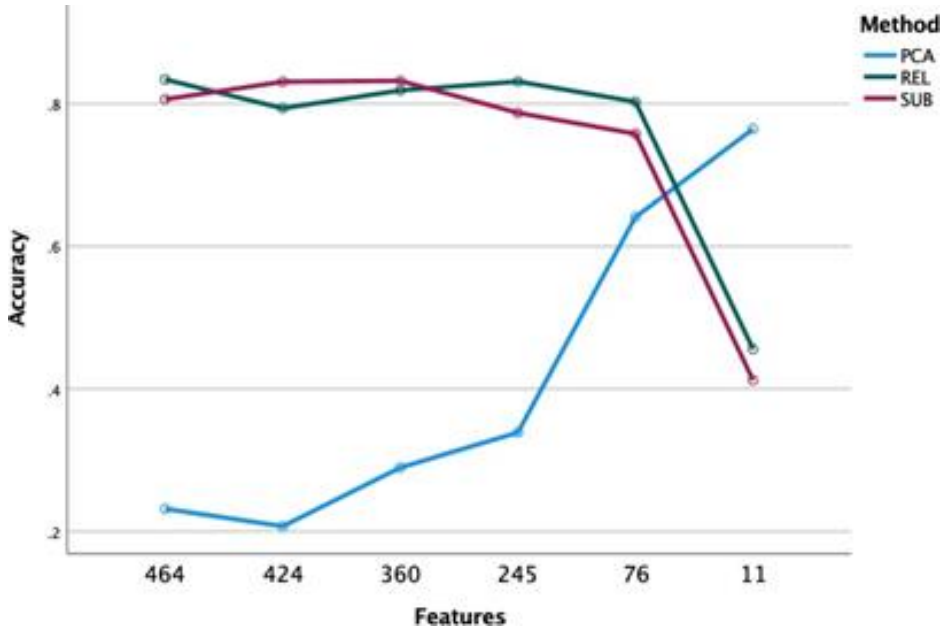

Figure 8. At lower feature levels, the PCA dimension reduction strategy performed best, at high levels of dimensionality performance of the PCA strategy falters. Results are pooled across all four classifiers.

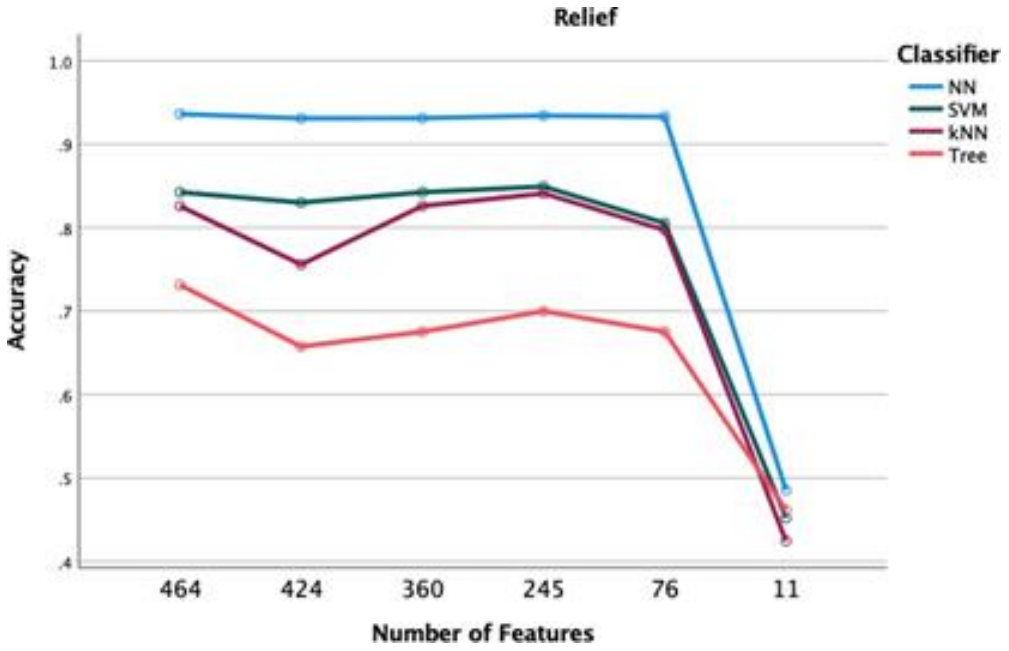

Figure 9. With dimension reduction by Relief, accuracy dropped below 76 features. The NN classifier performed best, the Tree classifier worst.

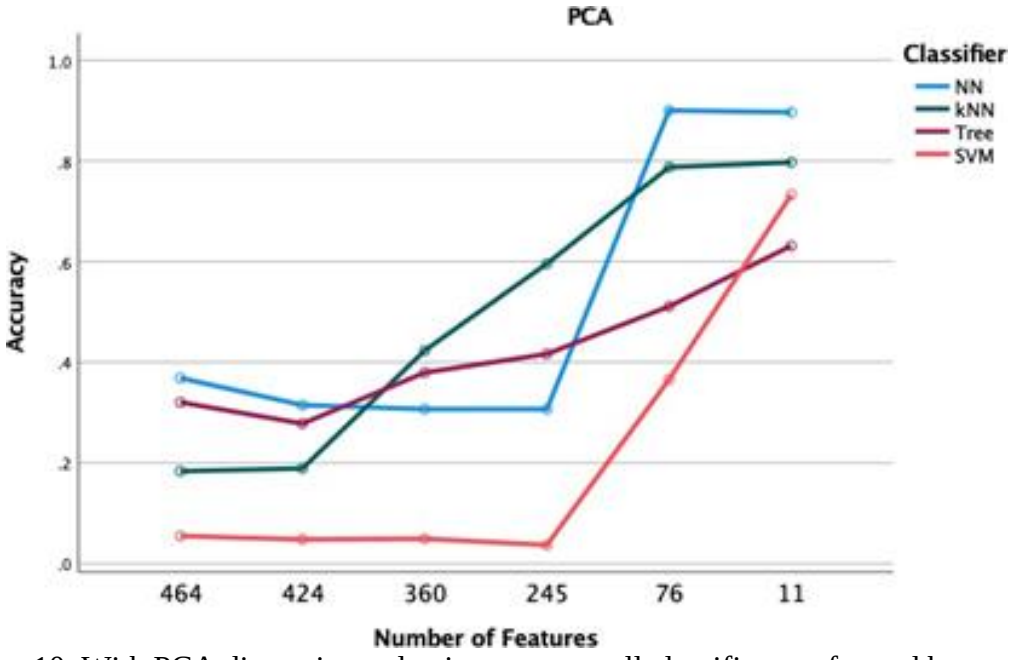

Figure 10. With PCA dimension reduction strategy, all classifiers performed better at 11 features than higher number of features. The NN classifier performed best and SVM performed worst with PCA strategy.

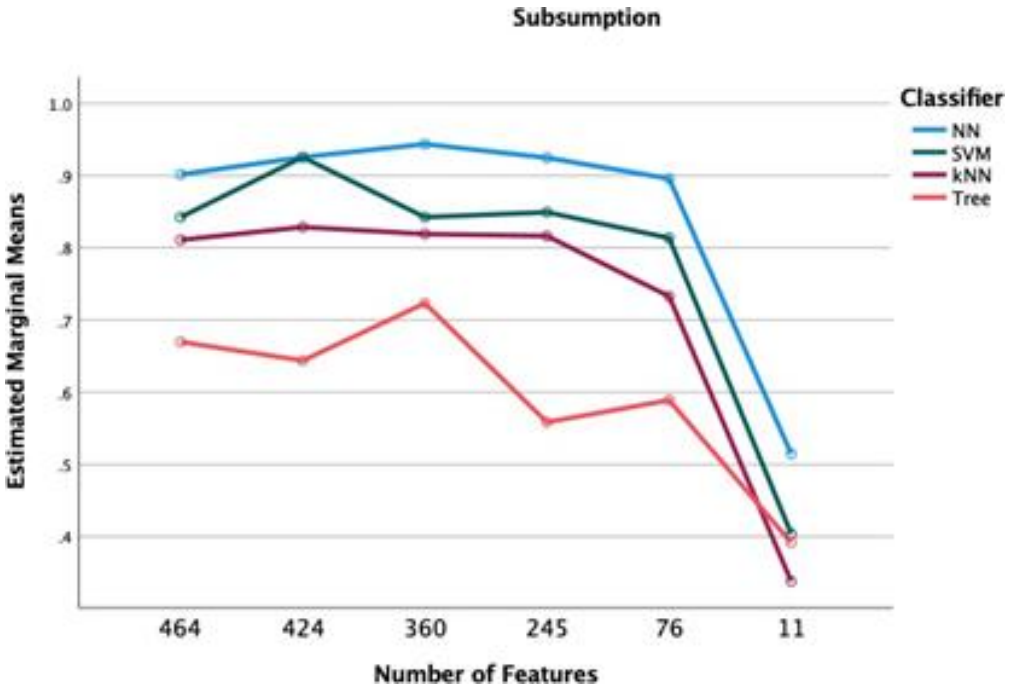

Figure 11. With dimension reduction by subsumption, accuracy begins to drop below 76 features. NN classifier performs best and Tree classifier worst with the subsumption strategy.

Across all models, the NN performed best for all dimension reduction strategies (Figure 6) on most features. We do see that the NN is less resistant to accuracy loss as the number of features increases for PCA compared to other models (Figure 10), which contributes to the lowered average accuracy at 245 features across all strategies (Figure 7). The fine tree classifier had the worst accuracy for all feature reduction strategies at any number of features.

Intuitively, one would expect that increasing the number of features would result in increased accuracy for any given model. What we have observed, however, is that for subsumption and relief strategies, there is little difference in accuracy between the greatest number of features (464) and the second smallest number of features (76) for our best performing models (NN, SVM). There is also little difference in model accuracy

between these strategies as well (Figure 9, 11). PCA dimension reduction strategies behave unintuitively, with smaller numbers of features being associated with greater accuracy (Figure 10).

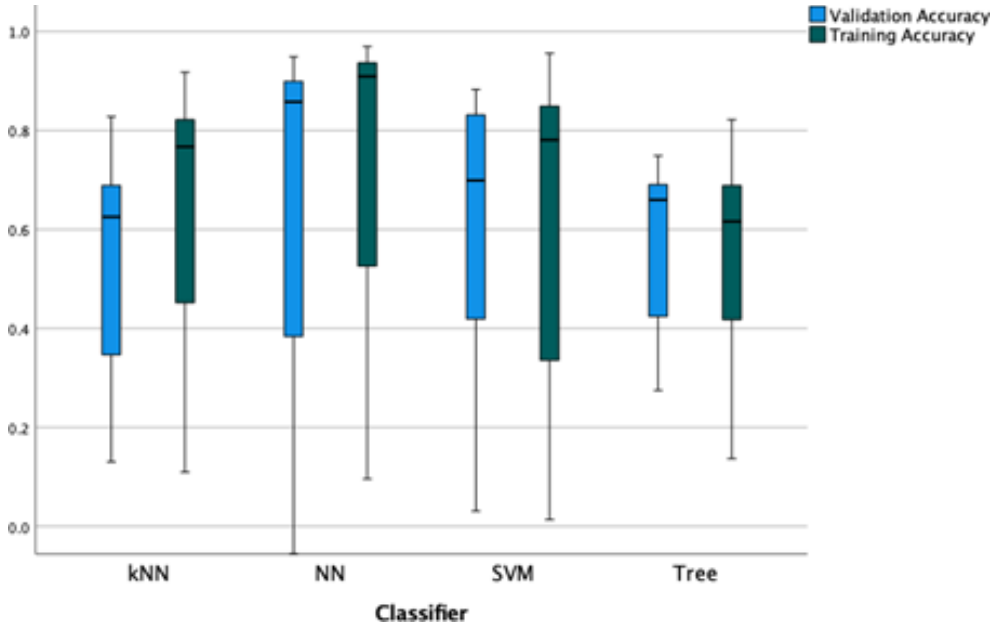

Figure 12. Test accuracy and Validation accuracy by classifier across all dimension reduction strategies and all feature levels.

Lastly, validation accuracy was assessed across all classifiers, features, and strategies. The NN maintained the highest median average accuracy for both the validation and test sets while the tree maintained the lowest median average accuracy for the test set and the kNN had the lowest 5-fold cross validation accuracy (Figure 12). The highest validation accuracy was maintained at 76 features, the highest test set accuracy at 424 features (Figure 13). Lastly, subsumption and relief both maintained higher validation and test set accuracy than PCA, and did not differ significantly from each other

(Figure 14). This does indicate that utilizing PCA may result in losses to model generalizability. Further investigation is needed to validate these results.

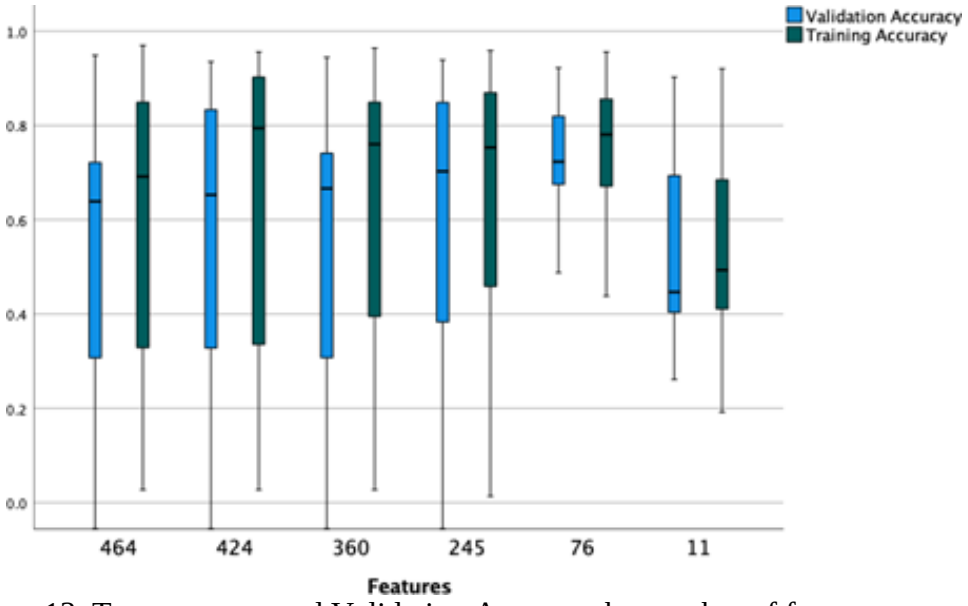

Figure 13. Test accuracy and Validation Accuracy by number of features across all classifiers and all dimension reduction strategies

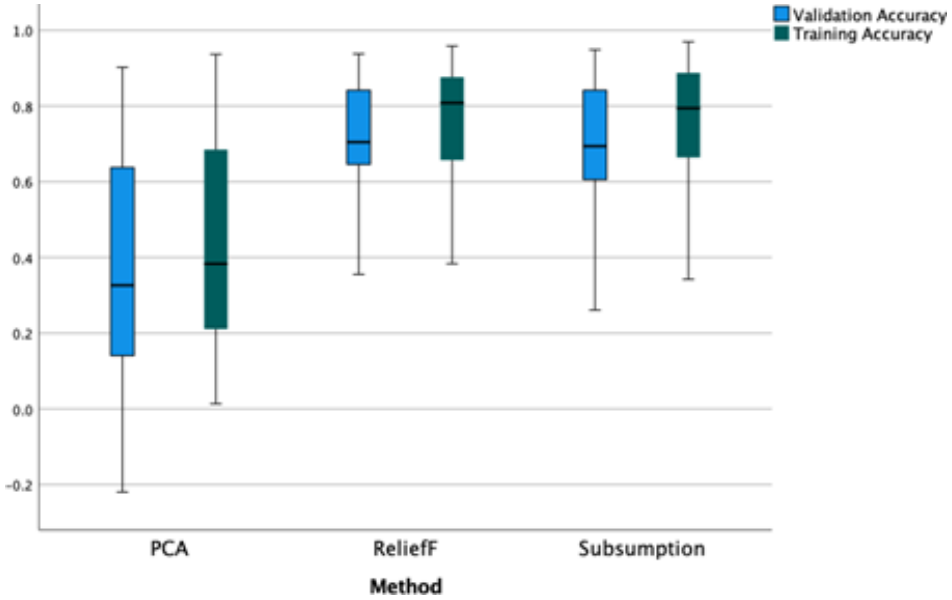

Figure 14. Test accuracy and Validation accuracy by dimension reduction strategy across all classifiers and all feature levels

4. DISCUSSION AND CONCLUSIONS

The features of our dataset were the signs and symptoms of patients with neurological diseases. The labels of our dataset were disease diagnoses. All features were one-hot encoded. Like many disease datasets, our dataset was of high dimensionality (475 different signs and symptoms) for 364 cases (Table 1). The classification task was to assign one of 20 different diagnoses to each of the 364 cases based on underlying features. High dimensionality poses difficulties for machine learning applications because of higher computational costs and the risk of including redundant or irrelevant features into the model.

The features of our dataset were derived from a subsumptive containment hierarchy [23]. In a subsumptive containment hierarchy, more specific concepts are subsumed by more general concepts. We used subsumption successively to reduce the dimensionality of our dataset from 474 dimensions to 11 dimensions. Each successive application of subsumption reduced dimensionality of the dataset and substituted a more general concepts for a more specific concept.

Several observations were notable:

- For all classifiers, the PCA dimension strategy worked best at lower levels of dimensionality (Figure 10 and Figure 8). Performance was best at 11 features and began dropping at 76 features for Tree and SVM and at 245 features for NN and kNN.
- Classification accuracy using ReliefF (Figure 9) and subsumption (Figure 11) did not fall until features were reduced below 76 features. For all

classifiers accuracy was lower for subsumption and ReliefF than PCA at the 11-feature level.

- Test accuracy and Validation accuracy was comparable across all experiments (Figures 12-14) suggesting that classification models were relatively robust.
- When averaged across all dimension reduction strategies, all classifiers performed best at 76 features (Figure 7).
- When averaged across all dimension reduction strategies, the NN classifier outperformed the SVM, kNN, and Tree classifiers (Figure 6).

The goal of dimension reduction methods for high dimension datasets is to find the minimal subset of features that maintains classifier accuracy and retains predicted class sizes reflective of the class sizes in the ground truth dataset upon retraining [33]–[35]. Two commonly used strategies to reduce dataset dimensionality include feature selection and feature extraction. Feature selection (filter methods, wrapper methods) emphasize algorithms that reduce the number of features into the smallest subset that accurately predict class membership [33]–[35]. Feature extraction methods (principal components, linear discriminant analysis, etc.) emphasize methods for collapsing many features into a smaller number of highly predictive features. The use of subsumption to collapse features into a smaller number of features bears more resemblance to a feature extraction strategy than a feature selection strategy. The use of knowledge embedded in a hierarchical ontology has been suggested by others as a dimension reduction strategy [36].

This work has important limitations. First, the dataset was small and future testing utilizing a larger dataset will be advantageous. Second, due to asymmetries in the depth of the ontology, the subsumption strategy only yielded six different levels of dimension reduction (464, 424, 360, 245, 76, and 11 features). To make comparisons fairly, we were limited to those dimensions by the subsumption strategy. We did not evaluate the performance of Relief or PCA at other levels of dimensionality, although those strategies could have created additional datasets of different dimensionality. Other studies have found that when different feature reduction strategies are compared classifier performance depends on the nature of the dataset, the classifer utilized, as well as the feature reduction algorithm [37]. Lastly, additional investigation into fine-tuning parameters for the various machine learning algorithms would have undoubtedly improved various architectures (particularly kNN, as the coarse kNN looked at the 100 closest neighbors of a dataset of 364, or coarse tree, which was limited to 4 splits).

REFERENCES

[1] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *Journal of the American Medical Informatics Association,* vol. 25, no. 10, pp. 1419–1428, 2018.

[2] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[3] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G.Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature medicine,* vol. 25, no. 1, pp. 24– 29, 2019.

[4] S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Ayme´, G. Baynam, S. M. Bello, C. F. Boerkoel, K. M. Boycott et al., "The human phenotype ontology in 2017," *Nucleic acids research*, vol. 45, no. D1, pp. D865–D876, 2017.

[5] S. Köhler, N. C. Øien, O. J. Buske, T. Groza, J. O. Jacobsen, C. McNamara, N. Vasilevsky, L. C. Carmody, J. Gourdine, M. Gargano et al., "Encodingclinical data with the human phenotype ontology for computational differential diagnostics," *Current protocols in human genetics*, vol. 103, no. 1, p. e92, 2019.

[6] T. Groza, S. Köhler, D. Moldenhauer, N. Vasilevsky, G. Baynam, T. Zemojtel, L. M. Schriml, W. A. Kibbe, P. N. Schofield, T. Beck et al., "The human phenotype ontology: semantic unification of common and rare disease," *The American Journal of Human Genetics*, vol. 97, no. 1, pp. 111–124, 2015.

[7] The National Center for Biomedical Ontology, "The human phenotype ontology," https://bioportal.bioontology.org/ontologies/HP, 2021, uploaded: 2020-12-07.

[8] M. Kuhn, K. Johnson et al., Applied predictive modeling. Springer, 2013, vol. 26.

[9] M. Kuhn and K. Johnson, Feature engineering and selection: A practical approach for predictive models. CRC Press, 2019.

[10] S. Visalakshi and V. Radha, "A literature review of feature selection techniques and applications: Review of feature selection in data mining," in 2014 *IEEE International Conference on Computational Intelligence and Computing Research*. IEEE, 2014, pp. 1–6.

[11] J. Demŝar, T. Curk, A. Erjavec,Črt Gorup, T. Hočevar, M. Milutinovič, M. Mŏzina, M. Polajnar, M. Toplak, A. Starič, M.Štajdohar, L. Umek, L.Žagar, J.Žbontar, M.Žitnik, and B. Zupan, "Orange: Data mining toolbox in python,"*Journal of Machine Learning Research*, vol. 14, pp.2349–2353, 2013. [Online]. Available: http://jmlr.org/papers/v14/demsar13a.html

[12] Bhatia, K.P.and Erro, R. and Stamelou, M., *Case Studies in Movement Disorders*, ser. Cambridge Medicine. Cambridge University Press, 2017.

[13] J. H. Noseworthy, *Fifty Neurologic Cases from Mayo Clinic*. Oxford University Press, 2004.

[14] S. Gauthier and P. Rosa-Neto, *Case Studies in Dementia: Volume 1: Common and Uncommon Presentations*, ser. Case Studies in Neurology. Cambridge University Press, 2011.

[15] H. Blumenfeld, *Neuroanatomy Through Clinical Cases*, ser. Blumenfeld, Neuroanatomy Through Clinical Cases. Sinauer Associates, 2010.

[16] J. Howard and A. Singh, *Neurology Image-Based Clinical Review*. Springer Publishing Company, 2016.

[17] E. C. Toy, E. Simpson, and R. Tintner, *Case Files Neurology, Second Edition*, ser. LANGE Case Files. McGraw Hill, 2012.

[18] S. A. Liveson, *Peripheral neurology: case studies*. Oxford University Press, 2000.

[19] S. Hauser, H. Weiner, and L. Levitt, *Case Studies in Clinical Neurology for the House Officer*. Williams & Wilkins, 1986.

[20] T. Solomon, B. Michael, A. Miller, and R. Kneen, *Case Studies in Neurological Infection*, ser. Cambridge medicine. Cambridge University Press, 2019.

[21] S. Waxman, *Clinical Neuroanatomy, 26th Edition*, ser. A Lange medical book. McGraw-Hill Education, 2009.

[22] S. Pendlebury, P. Anslow, and P. Rothwell, *Neurological Case Histories: Case Histories in Acute Neurology and the Neurology of General Medicine*, ser. Oxford Case Histories. Oxford University Press, 2007.

[23] D. B. Hier and S. U. Brint, "A neuro-ontology for the neurological examination," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–9, 2020.

[24] D. B. Hier, J. Kopel, S. U. Brint, D. C. Wunsch, G. R. Olbricht, S. Azizi, and B. Allen, "Evaluation of standard and semantically-augmented distance metrics for neurology patients," *BMC Medical Informatics and Decision Making*, vol. 20, no. 1, pp. 1–15, 2020.

[25] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[26] M. Ringnér, "What is principal component analysis?" *Nature biotechnology*, vol. 26, no. 3, pp. 303–304, 2008.

[27] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[28] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in *Aaai*, vol. 2, 1992, pp. 129–134.

[29] I. Kononenko, E. Sˇimec, and M. RobnikSˇikonja, "Overcoming the myopia of inductive learning algorithms with relieff," *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997.

[30] A. J. Smola and B. Scho¨lkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[31] C. Kingsford and S. L. Salzberg, “What are decision trees?” *Nature biotechnology*, vol. 26, no. 9, pp. 1011–1013, 2008.

[32] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, “A tutorial on the cross-entropy method,” *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[33] M. Dash and H. Liu, “Feature selection for classification,” *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[34] J. Tang, S. Alelyani, and H. Liu, “Feature selection for classification: A review,” *Data classification: Algorithms and applications*, p. 37, 2014.

[35] D. Koller and M. Sahami, “Toward optimal feature selection,” Stanford InfoLab, Tech. Rep., 1996.

[36] D. C. Corrales, E. Lasso, A. Ledezma, and J. C. Corrales, “Feature selection for classification tasks: Expert knowledge or traditional methods?” *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 5, pp. 2825–2835, 2018.

[37] A. Janecek, W. Gansterer, M. Demel, and G. Ecker, “On the relationship between feature selection and classification accuracy,” in *New challenges for feature selection in data mining and knowledge discovery*, 2008, pp. 90–105.

SECTION

2. CONCLUSIONS

In this research, feature reduction techniques for use with machine learning approaches have been presented in medical applications. Specifically, research has explored the diagnosis of disease from patient signs and symptoms has been conducted on numerous machine learning algorithms, including a neural network, various support vector machines, various k-Nearest Neighbor algorithms, and various classification trees. In addition, a proof-of-concept application of utilizing reinforcement learning was discussed, that if further investigated and improved, could offer a method of reliably creating models of various physiological processes, provided an available mathematical foundation. For this work, methods for reducing dimensionality has been investigated in complex medical datasets. Our experimental studies indicate that various feature selection methods can be implemented in our data and still preserve algorithm accuracy. Further research must be done to validate our results. In addition, our preliminary reinforcement learning model shows potential for developing complex models of the interactions of varied organ systems in biological processes and regulation.

BIBLIOGRAPHY

[1] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," *Journal of the American Medical Informatics Association*, vol. 25, no. 10, pp. 1419–1428, 2018.

[2] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learn- ing for healthcare: review, opportunities and challenges," *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236–1246, 2018.

[3] Hamm LL, Nakhoul N, Hering-Smith KS. Acid-Base Homeostasis. Clin J Am Soc Nephrol. 2015;10(12):2232-2242. doi:10.2215/CJN.07400715

[4] M. Kuhn, K. Johnson et al., *Applied predictive modeling*. Springer, 2013, vol. 26.

[5] M. Kuhn and K. Johnson, *Feature engineering and selection: A practical approach for predictive models*. CRC Press, 2019.

VITA

Donald Coolidge Wunsch III was a graduate student at Missouri University of Science and Technology (Missouri S&T). He was a dual-enrolled undergraduate/graduate student and received his Bachelor of Science in Computer Engineering from Missouri S&T in May of 2021. Upon graduation, he accepted an Intramural Research Training Award from the National Institutes of Health (NIH) and continued to work towards his graduate degree. His interests in computer engineering and medicine overlapped in the field of machine learning, particularly, bioinformatics. This interest motivated him to pursue a master's degree at Missouri S&T. He worked heavily with Dr. Daniel Hier to find applications of machine learning in medicine. He received a Master of Science in Computer Engineering from Missouri S&T in December 2021.