August 2021

Prediction of Concurrent Hypertensive Disorders in Pregnancy and Gestational Diabetes Mellitus Using Machine Learning Techniques

Mary Ejiwale
University of Wisconsin-Milwaukee

PREDICTION OF CONCURRENT HYPERTENSIVE DISORDERS IN PREGNANCY AND GESTATIONAL DIABETES MELLITUS USING MACHINE LEARNING TECHNIQUES

by

Mary O. Ejiwale

A Dissertation Submitted in

Partial Fulfilment of the

Requirements for the Degree of

Doctor of Philosophy

in Biomedical and Health Informatics

at

The University of Wisconsin-Milwaukee

August 2021

ABSTRACT

PREDICTION OF CONCURRENT HYPERTENSIVE DISORDERS IN PREGNANCY AND GESTATIONAL DIABETES MELLITUS USING MACHINE LEARNING TECHNIQUES

by

Mary O. Ejiwale

The University of Wisconsin-Milwaukee, 2021
Under the Supervision of Professor: Susan McRoy, PhD

Gestational diabetes mellitus and hypertensive disorders in pregnancy are serious maternal health conditions with immediate and lifelong mother-child health consequences. These obstetric pathologies have been widely investigated, but mostly in silos, while studies focusing on their simultaneous occurrence rarely exist. This is especially the case in the machine learning domain. This retrospective study sought to investigate, construct, evaluate, compare, and isolate a supervised machine learning predictive model for the binary classification of co-occurring hypertensive disorders in pregnancy and gestational diabetes mellitus in a cohort of otherwise healthy pregnant women. To accomplish the stated aims, this study analyzed a sub-sample (n=4624, n_features=38) of a labelled maternal perinatal dataset (n=9967, n_fields=79) collected by the PeriData.Net® database from a participating community hospital in Southeast Wisconsin between 2013 and 2018. The datasets were named, “WiseSample” and “WiseSubset” respectively in this research. Thirty-three models were constructed with the following six supervised machine learning algorithms: Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, StackingClassifier, and KerasClassifier (which is a deep learning

classification algorithm). All the algorithms were evaluated using the StratifiedKfold cross-validation (k=10) method. The Synthetic Minority Oversampling Technique was applied to the training data to resolve the class imbalance that was noted in the sub-sample at the preprocessing phase. Multiple feature selection techniques were explored to identify the best predictors of concurrent hypertensive disorders in pregnancy and gestational diabetes mellitus. Model performance quality was quantitatively evaluated and compared using accuracy, F1, precision, recall, and the area under the receiver operating characteristic curve as metrics.

Support Vector Machine objectively emerged as the most generalizable model for identifying the gravidae in WiseSubset who may develop concurrent hypertensive disorders in pregnancy and gestational diabetes mellitus. The model obtained a recall score of 100.00% (mean), with 9 predictors extracted by the recursive feature elimination with cross-validation with random forest. Finding from this study show that using readily available routine prenatal attributes, appropriate machine learning methods can reliably predict the co-existence of hypertensive disorders in pregnancy and gestational diabetes mellitus. Six of the nine most predictive factors of the comorbidity were also in the top 6 selections of at least one other feature selection method examined. The six predictors are ***healthy weight prepregnancy BMI, mother's educational status, husband's educational status, husband's occupation in one year before the current pregnancy, mother's blood group, and mother's age range between 34 and 44 years.*** Insight from this analysis would support clinical decision making of the obstetric experts when they are caring for 1.) the primigravidas since they would have no past obstetric history that could prompt their care providers for related feto-maternal medical surveillance; and 2.) the multigravidas with no previous pregnancy history that is suggestive of hypertensive disorders in

pregnancy or gestational diabetes mellitus. Ultimately, the artificial-intelligence-backed tool designed in this research would likely improve maternal-child care quality outcomes.

Keywords: Gestational Diabetes Mellitus, Hypertensive Disorders of Pregnancy, Supervised Machine Learning, Deep Learning, Comorbidity, Concurrence.

This Dissertation is Dedicated

to my

Parents.

Oluwa Seun Fun Ohun Gbogbo!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Acronyms

ACKNOWLEDGEMENTS

I am indebted to my academic advisor Dr. Susan McRoy, who has been providing me guidance since 2019 when I had to switch advisor because my then primary advisor; Late Dr. Patrick (of blessed memory), was working towards his retirement. My unreserved gratitude goes to you, Dr. McRoy; your time, wealth of knowledge shared with me, and the passion displayed for my success are invaluable. Thank you very much! As for Dr. Patrick, his usual phrase; "let your good enough be good enough", resonates with me, and I find it useful in all areas, may his gentle soul rest in perfect peace!

To my doctoral committee members: Dr. Jake Luo, Dr. A Palatnik, Dr. Yetunde Folajimi; and Dr. Teresa Johnson, thanks so much for your feedback. You also demonstrated you were all out for my success, and I thank you dearly. I specially acknowledge Dr. Teresa Johnson, who is the Principal Investigator of this research through data procurement. I must recognize the informal support of Drs. Christine Cheng, and Wilkistar Otieno. These phenomenal women inspire me in many ways, thanks for being my good mentors!

My appreciation is incomplete if I fail to recognize the opportunity given me to be part of the 500 STARS program of the Clinical and Translational Science Institute (CTSI) of the Southeast Wisconsin as an intern (Summer 2017/2019) in the Biomedical Informatics department of Medical College of Wisconsin (MCW). Thanks to Ms. Memory Bacon, Ms. Chamia Gary, and Dr. Doriel Ward who is the Executive Director of the CTSI Administration. The same magnitude of gratitude goes to Mr. Bradley Taylor; the director of the Biomedical Informatics department, MCW; and two of his staff members; Mr. Wes Rood and Mr. Elias Devoe who individually were

my preceptors during the internship. The practical experience that I gained that time strengthened my interest in data science, and it boosted my dexterity tremendously.

For the GAANN fellowship that funded my PhD program (for the most part), Long Live the United States of America! I also thank the Computer Science department, UWM, for the teaching assistantship offers, the experience garnered goes a very long way! Also notable is the assistance of the dissertation formatting office of the graduate school, UWM, for ensuring this document is in the expected format, thanks a lot. My appreciation extends to Ms. Dunlap Coleen of the Financial Aid Office, UWM, and Ms. Chantelle Hoover-Mallette for her administrative services. Thanks to my collaborative friends at UWM; Sammie Omranian, Claudia Gallegos, Xiaoyu Liu, and Ling Thong, you guys are awesome (come snow, come sunshine)! Many other professors, teachers, and classmates have positively influenced me in one way or the other, and I value their effort as well.

Tremendous appreciation goes to my children for their **LOVE,** help, understanding, moral support, and endurance; all these motivated my perseverance, this degree is for you three!!
I equally express my sincere gratitude to my siblings for their encouragement in all ramifications, we all did it together!

Above all, to God alone be the glory----- !!

1. INTRODUCTION

1.1. OVERVIEW OF THE PROBLEM

1.1.1. OVERVIEW OF GESTATIONAL DIABETES MELLITUS AND HYPERTENSIVE DISORDERS IN PREGNANCY

Gestational diabetes mellitus (GDM) and hypertensive disorders in pregnancy (HDP) are two principal medical conditions that complicate pregnancy. They affect pregnant women of both the developed and developing countries. These maternal health problems can occur alone or simultaneously. Although both GDM and HDP have been widely studied, they have been examined individually mostly, while their comorbidity is rarely investigated, especially with machine learning (ML) methods. This research utilized five evidence-based standard supervised machine learning (SML) algorithms, and a classifier from the deep learning (DL) sub-field of SML for studying the co-existence of HDP and GDM (GDHP) affecting some otherwise healthy pregnant women. The standard SML algorithms tested were Logistic Regression, Random Forest, Decision Tree, Support Vector Machine, and StackingClassifier. Also, the DL algorithm explored is the KerasClassifier.

Hypertensive disorders in pregnancy are a group of maternal health conditions previously known as pregnancy-induced hypertension. It is characterized by a systolic blood pressure ≥140 mmHg and/or a diastolic blood pressure at ≥90 mmHg taken at least on two occasions of 4 hours apart in previously normotensive women [1] at or after 20 weeks of gestation. There are five main variants of HDP: gestational hypertension pre-eclampsia (with and without severe features), eclampsia, and HELLP (Hemolysis, Elevated Liver Enzymes and Low Platelet) syndrome. All were present in the data analyzed in this research, and they were investigated collectively as

HDP because studies have shown their risk factors are largely similar. Pregnancy is normally associated with some hemodynamic changes and an impairment in such bodily modifications may result in HDP [2]. Some of the maternal and child impacts of HDP are an increased risk for future chronic hypertension [3] and cardiovascular disease for both mother [4] and child [5]. Chronic hypertension (which exists or is diagnosed before 20th week of pregnancy) and pre-pregnancy/pre-existing diabetes were excluded from this study. This exclusion is reasonable since such medical histories are already a clear signal for the provider to institute "high risk" pregnancy management. Also, postpartum hypertension (PHTN) is outside the scope of this study, and a concurrence between PHTN and GDM is not meaningful since GDM is not a postpartum diagnosis, and no data related to postpartum glucose test is in the analyzed dataset.

Gestational diabetes mellitus is the most common medical complication of pregnancy. Insulin resistance and the accompanying compensatory hyperinsulinemia by the pancreatic β-cells are some of the physiologic changes of pregnancy. However, some women experience an imbalance in these normal processes, leading to gestational diabetes. GDM is a hyperglycemic condition that starts newly (or is first diagnosed) at late pregnancy ≥24th week in an otherwise euglycemic woman. This pregnancy complication is characterized by glucose intolerance, insulin resistance, and hyperglycemia [5]. In terms of screening/evaluation there is no universal assessment guideline currently for GDM [6]. Expectant mothers are usually screened for this health problem between 24 and 28 weeks of pregnancy, using Oral Glucose Tolerance Test (OGTT) to evaluate the efficiency of the body to metabolize glucose. Like its HDP counterpart, GDM has a plethora of deleterious immediate and future maternal and child effects. These include an increased risk for diabetes (both mother and offspring), early faster puberty for the child [7].

1.1.2. OVERVIEW OF MATERNAL METABOLIC SYNDROME

Metabolic syndrome (MetSyn), previously known as Syndrome X (or insulin resistance syndrome), has a significant relationship with HDP and GDM. A shared maternal impact of HDP and GDM is their metabolic sequalae [8], and MetSyn may predict HDP [9], and/or gestational diabetes [10]. Literature also indicates women's sex roles (parturition-specific factors) such as pregnancy, parity, lactation, contraception, and infertility treatment; influence the risk of MMetSyn [11], [12], [13], [14]. The condition is characterized by a cluster of five major risk factors, three of which are closely related to HDP and GDM. The established risk factors for MetSyn are high body mass index; insulin resistance/ high blood glucose; high level of low-density lipoprotein, low level of high-density lipoprotein [dyslipidemia]; and high blood pressure [15]. MetSyn is a grievous global health challenge with no regard for age, gender, or ethnicity. The syndrome is defined by some national and international health bodies, including the World Health Organization [16], International Diabetes Federation [17], and the National Cholesterol Education Program Adult Treatment Panel III 2001 [18]. Even though these organizations differ slightly in their full definitions of MetSyn, they are unanimous in including a measure of central obesity, glucose metabolism, and blood pressure in their characterizations of the syndrome. MetSyn has had increasingly wide scientific attention, but the general population is commonly targeted [19], [20], [21]. This focus may have contributed to the lack of maternal metabolic syndrome (MMetSyn) as an obstetric diagnosis, and thus a lack of an associated International Classification of Diseases 9/10 (ICD Version 9/10) code, despite the association of MetSyn with GDM, HDP and parturition. These issues about MetSyn of the pregnant women have also resulted in a lack of easily accessible structured information about the health problem in electronic health records of pregnant women that might allow one to easily extract related

dataset. This makes a direct, manual cohort analysis impractical. Instead, the research described here considers the relationship between GDM, HDP and MetSyn, using a method that creatively serves as a proxy for a retrospective cohort analysis where machine learning methods are applied to predict maternal metabolic syndrome. Moreover, this innovative proxy study has a utility, since GDM and HDP are well recognized diseases in the obstetric community than maternal MetSyn, building a model for their comorbidity is more valuable to the study population and their providers, than modelling MMetSyn that is yet to be established as a diagnosis. This type of study is particularly useful with sophisticated methodology like machine learning; the kind utilized in this research.

1.2. OVERVIEW OF THE DATASET

The PeriData.Net® database is a perinatal clinical data repository. It consists of individual-level patient-identifiable data collected from the parents and the clinical record at participating Wisconsin (WI) birth hospitals. The database was originally developed in the mid-2000s for collecting Birth Certificate Data required by the Department of Health Statistics (DHS), with additional information as requested by participating hospitals, the Wisconsin Association of Perinatal Care (WAPC), and other stakeholders. Ultimately, DHS chose another program to collect the data, but most of the birth hospitals in the state continue to use Peridata.net®[18] because there are valuable reporting formats provided by Ancilla, LLC; a company that is responsible for the data collection and storage.

This research is being conducted using a dataset from a single hospital in a small urban community in the Southeast WI for the period of 1/1/2015 to 2018. This sample contains 9962 instances with 79 fields, while its extract has 4624 instances with 38 fields. The omitted fields included: 1.) primary fields that were combined during data extraction; 2.) fields with high

number of missing values (e.g., intrauterine growth retardation; 3.) fields with low frequency count values that could not benefit from merging (father's education for instance); 4.) fields about/clearly suggestive of GDM/HDP-in the past obstetric history ("macrosomia_PreviousPreg" for example); and 5.) redundant fields (for example, husband's reported age and husband's calculated age fields have nearly the same number of actual values). Our research protocol included two types of limited Protected Health Information (PHI): "Year of last birth" and "Date of first prenatal visit" fields, which were added because the contained information that could be used to derive some known risk factors of GDM and HDP. There were three types of values in the sample- actual, missing, and placeholder. Also, the dataset format is heterogeneous, having a disparate mix of numeric, and non-numeric types that could be mapped onto numeric format during preprocessing. Some numeric fields in the dataset also have values that could be manipulated into new fields to enable the identification of the specific/range of values associated with GDHP; prepregnancy Body Mass Index (BMI) was one of such primary fields. Additionally, the sub-sample dataset is significantly imbalanced for the outcome values.

1.3. OVERVIEW OF METHOD

A sample of PeriData.Net® database was analyzed in a retrospective cohort study that examined a representative set of supervised machine learning algorithms for constructing several binary classification models for identifying the GDHP-at-risk gravidae. The six classifier training algorithms examined in this study were Logistic Regression (LReg), Random Forest (RF), Decision Tree (DTree), Support Vector Machine (SVM), StackingClassifier (Stack), and KerasClassifier (Keras), which is a deep learning method. During preliminary training and testing, we generated several alternative sets of features using four well-known feature selection techniques (FSTs). The following FSTs were applied 1.) Recursive Feature Elimination (RFE),

2.) Recursive Feature Elimination with Cross-Validation (RFECV), also known as ensemble feature selection approach; 3.) Genetic Algorithm, and 4.) Pearson Correlation Coefficient and Variance Inflation Factor Analyses (PCC-VIFA).

The impact of the features and the strategy that selected them were assessed through the performance of models constructed with them. The quantitative quality metrics utilized for these assessments were accuracy, F1, precision, recall, and the area under the receiver operating characteristic curve (AUC), including a plot of the receiver operating characteristic curve (ROC).

1.4. OBJECTIVES, RESEARCH QUESTIONS, AND SPECIFIC AIMS

1.4.1. GENERAL OBJECTIVES OF THE STUDY

This study sought to construct, evaluate, and compare multiple supervised machine learning (standard and deep) models for pregnant women who are at risk for concurrent gestational diabetes mellitus and hypertensive disorders in pregnancy. The motivation behind the plan to explore and compare multiple classification models is the idea that no single model performs optimally across all problems; a phenomenon known as *"No Free Lunch theorem"* in the machine learning domain. The problem of predicting the risk for maternal MetSyn was mapped to a problem of developing a classifier to map the multi-facetted data of the gravidae to a binary class outcome of Yes_GDHP or No_GDHP. For the analysis. we examined the data fields that already exist in the dataset and some other fields that we engineered from them. All the models were to be assessed quantitatively and compared for performance quality, using standard measures of accuracy, F1, precision, recall (sensitivity), and the area under the receiver operating characteristic curve (AUC) as metrics. The best performing model that generalized well to

unseen data in the identification of the GDHP-at-risk instances was to be isolated, and the following research questions were to be addressed:

1.4.2. RESEARCH QUESTIONS

1.4.2.1. RESEARCH QUESTION 1

Compared to genetic algorithm, does any of the feature selection techniques (Section 1.3) better identify the best input data for building an SML model for GDHP with the dataset in this analysis?

1.4.2.2. RESEARCH QUESTION 2

Could there be any GDHP model that can outperform the Keras model when assessed with recall on the dataset to be analyzed in this study?

1.4.2.3. RESEARCH QUESTION 3

Would the Synthetic Minority Oversampling Technique (SMOTE) algorithm effectively address the class imbalance problem that exists in the dataset?

These questions are answerable by the corresponding specific aims of the project (Section 1.4.3)

1.4.3. SPECIFIC AIMS OF THE STUDY

The specific aims are to 1.) assess the utility of the FSTs (Section 1.3) in identifying the most relevant risk factors for modeling GDHP with SML techniques; 2.) utilize and compare multiple SML algorithms (Section 1.3) in building GDHP models with the dataset; 3.) construct and compare various SML models (Section 1.3) before and after the application of SMOTE on the imbalanced dataset. These aims would lay the groundwork for assessing the feasibility of

automated analyses, and for establishing the potential benefit of conducting studies with larger datasets in the future.

2. LITERATURE REVIEW AND RELATED STUDIES

2.1. LITERATURE REVIEW

2.1.1. BACKGROUND OF THE OBSTETRIC HEALTH PROBLEMS

Gestational diabetes mellitus and hypertensive disorders of pregnancy are common medical conditions associated with pregnancy. The prevalence of HDP is 5.2-8.2% [22]. In a multinational, multi-site research conducted in 2010-2012; where 214,070 women of 106 communities in 7 low and middle-income nations were studied, 16% (55) of the 335 women that died had preeclampsia or eclampsia [23]. In Southeast Iran 31.9% of severe maternal morbidity near misses [24]. The Centers for Disease Control and Prevention (CDC) indicates DHP attributes to 6.6% of pregnancy and childbirth related deaths [25]. Gestational diabetes mellitus on the other hand, is estimated to constitute 16.0% of the global prevalence of hyperglycemia in pregnancy among women ages 20-49 (16.9%) [26]. Insulin resistance and the accompanying compensatory hyperinsulinemia by the pancreatic β-cells are some of the physiologic changes of pregnancy. However, an imbalance in these processes may result in a hyperglycemic condition that starts newly (or is first diagnosed) in late pregnancy ≥24^{th} week in the otherwise euglycemic pregnant women. GDM and HDP are individually consequential pregnancy-associated health conditions with immediate and lifelong maternal and fetal/child health impact. Mounting evidence is available in literature pointing to the immediate and future (sequelae) maternal and child health (MCH) effect of HDP and GDM.

2.1.1.1. MATERNAL AND CHILD IMPLICATION OF GESTATIONAL DIABETES MELLITUS OR HYPERTENSIVE DISORDERS OF PREGNANCY

Pregnant women experiencing HDP may have an increased risk for chronic hypertension 5 years postpartum [27], and cardiovascular disease for both mother [4], [28], and baby [5]. Post-

traumatic stress disorders [29]; salt hypersensitivity [30]; end-stage renal disease (increased risk by 5- to-12-folds) [31], [32]; and metabolic syndrome [33] , [8] are also linked to HDP or GDM. Infants born to pregnancy-related hypertensive or diabetic mothers are not spared of the short and long-term effect of these disorders. The child impact includes the following: An increased risk of cardiovascular disease [5]; and neuro-developmental impairment [34]; and future weight problem [35]; metabolic syndrome [36]. Meanwhile, MetSyn increases the chance of the boy child experiencing infertility in his adulthood [37]. Likewise, prematurity and birth weight problems are also linked to HDP [38] and GDM [39]. In a ripple effect, female premature babies/girls who had birth-weight problem have a high tendency of developing HDP or GDM when pregnant [40] perpetuating the cyclic nature of the disorder. Other impacts of HDP/GDM include neuro-developmental impairment [34], [41] and behavioral disorder [42].

2.1.1.2. COST IMPLICATION OF GDM OR HDP

Aside from the scores of maternal-fetal/child consequences of HDP and GDM, evidence abounds in the literature showing these pregnancy-related medical conditions are also uneconomical. HDP alone incurs additional care costs of $173 million annually [43], and the annual burden per case of GDM is approximately $5,800 [44].

2.1.1.3. RISK FACTORS OF GDM OR HDP

The unpleasant, and sometimes life-threatening, effect of GDM or HDP is heightened when both disorders co-exist, then, GDHP deserves additional scientific attention. Meanwhile, the exact causes of GDM and HDP individually or jointly, are unknown, but related studies suggest several factors explain them. It is worthwhile investigating such risk factors (and metabolic syndrome) to understand and identify the strongest predictors of GDHP. Among the social determinants of health influencing these pregnancy complications are smoking, alcoholism, and

socio-economic factors [45], maternal height [46], parity [47], [12], gravidity [12], maternal blood group [48], [49], [50], outdoor air/temperature [51], and genetics [52]. Research shows prepregnancy obesity [53] is fundamental to these metabolic-related obstetric health problems. Pre-conception weight/BMI is commonly based on the score obtained at the first prenatal care visit. Obesity is a body mass index (BMI) ≥30 kg / m^2, according to the CDC [54]. A substantial body of evidence demonstrates the determinants of health influencing of GDM and HDP; including their conjoint, is multifactorial, and while some risk factors are peculiar to the expectant mother, the unborn child, and his/her biological father; others have genetic, environmental, or obstetric components. Studies show assisted reproductive technology (ART) is also associated with gestational hypertension and preeclampsia [55]. This assertion is supported by a large systematic review of 47 related studies [56]. Additionally, a meta-analysis of twenty-six GDM studies with 120 million participants signifies mother's age is crucial to the development of GDM or HDP, but the study asserts that there is no agreement over the specific age/ age group associated with GDM [57]. The role, and the precise maternal age/age group that is linked with GDHP is, however, not yet established. Maternal height is a significant anthropometric measurement in hypertensive disorders of pregnancy [58], and short people have a higher predisposition to GDM [46]. It is unclear whether this factor influences the concurrence of both disorders. Central to these maternal health problems is the obesity epidemics [14]. Studies indicate that, in the U.S.; where metabolic syndrome surges [59], prepregnancy obesity deepens [12]. As shown in a recent analysis of the Wisconsin Interactive Statistics on Health (WISH) data of the 2011–2014 period, 27.8% of WI mothers are obese [60]. Some pregnant women, however, may not know their prepregnancy weight or BMI for reasons such as unintended pregnancy (>50.00% in the United States [61]), hence, the commonest approach is the adoption of the first prenatal visit values of these measurements. In a research that utilized animal model (rats), the results suggest exposure

to air pollutants such as nicotine (first/secondhand) contributes almost one-third (30%) of maternal obesity [62] and maternal metabolic conditions [14]. Other significant environmental components influencing HDP/GDM occurrence are Particulate Matter 7 (PM7) vehicular emits exposure [63], and urbanization [64]. Furthermore, there is evidence supporting the notion that vitamin D; also known as 25-hydroxyvitamin D (25(OH)D), has a bearing on pregnancy outcomes. The environment, genetics, and racial are a close- knit with this fat-soluble vitamin that enhances glucose tolerance. Vitamin D deficiency (hypovitaminosis D), which is defined as low serum 25-hydroxyvitamin D (25(OH) D) level [65], is prevalent worldwide [66], (among women particularly). Maternal hypovitaminosis D raises the susceptibility to GDM and HDP [67], [68]. This condition is, however, racially biased against the African Americans and their high level of melanin is linked to the insufficiency [66]. This racial group is the worst hit of preeclampsia [69] and maternal mortality in the United States [70]. Partner change is another unconventional factor influencing HDP; preeclampsia especially [55], [58]. Access to care and medical insurance [45]; and fetal gender [71] are some other dynamics influencing GDM and HDP. It is therefore evident that the impact of GDM and HDP is both extensive and expensive, and the coalesce of the disorders may escalate their individual devastating consequences. Also, the challenges confronting the existing risk identification approaches are complicated. These issues affirm the relevance of our research as it is of utmost importance to understand this serious comorbidity from supervised machine learning perspective.

2.1.1.4. THE NEED FOR MACHINE LEARNING PREDICTIVE MODELS FOR GDHP

The cause(s) of gestational diabetes mellitus and hypertensive disorders in pregnancy remains largely elusive, and their prevention is a challenge to the obstetric community [72]. There is lack of direct laboratory test or a particular screening method designed for identifying the GDHP-at-

risk pregnant women; the test (including results) for each of the disease components of GDHP may also reside across multiple silos screening; and the tests may have been done at different times or places. Such disjoint screening is not only inconvenient for pregnant women, but also resource-consuming (patient time, care cost; and hospital economy) for both patients and providers. Some of the individual screening methods may still have limitations, as Gaillard et al. (2018) [73] indicated about the African American women and metabolic syndrome screening. Additionally, late GDM screening [74], and the variations of such screening [6] pose challenges to the disease. Lack of agreement over laboratory tests threatens the quality of the procedure, and it may endanger perinatal outcomes. Another issue with GDHP risk assessment currently is that early clinical diagnosis is hampered by the mid (HDP) to late (GDM) manifestation of the clinical symptoms of each of its disease components. Lastly, convoluted interactions exist among the many factors associated with the comorbidity of HDP and GDM, and these interactions are not only difficult to determine, but also too burdensome for most providers' limited time. Therefore, there is a great need for GDHP risk factor assessment automation through the machine-aided approach that this study develops. Such a tool is scalable, and it would ameliorate the discussed challenges with fewer resources. It would also be more convenient; requiring no patient effort, and generate results instantaneously at the point of care, all in a feto-maternal non-invasive manner. Moreover, because this study would identify which group of the gravidae have the greatest risk of GDHP, and what the best predictors are, it offers insight into the care providers and the stakeholders towards targeted and strategic program planning (prevention and early intervention). This would ultimately save the associated cost from increasing morbidity and mortality to the care of women and their infants. As advantageous as risk identification of GDHP might be to perinatal outcomes, the comorbidity is rarely studied, relative to research

investigating GDM and GDP individually, especially in the machine learning domain, therefore, this research also adds to the body of knowledge.

Finding from this work would support clinical decision making of obstetric experts to identify pregnant women who have a high risk for GDHP and thus might need more specialized care management. The model will particularly be useful to them when caring for 1.) the primigravidas since such women are carrying their first pregnancy, thereby, have no previous obstetric history that could prompt their care providers for maternal and child medical surveillance; and 2.) the experienced pregnant mothers (multigravidas) with no known obstetric history related to GDM, HDP or GDHP. Therefore, the model will likely boost maternal and child care quality outcomes; and ultimately be lifesaving for pregnant women and their unborn babies. Equally, the result of the proposed study will potentially promote an effective targeted preventive/interventional patient care plan; and proper resource distribution to the most needful because our model will additionally identify the specific non-causal (risk) factors that best explain GDHP in a non-invasive manner.

2.1.2. BACKGROUND OF METHOD

The methods will examine the effectiveness of alternative automated methods for training predictive models with existing data where the data elements correspond to the aggregated clinical history of an individual patient, represented as a set of fields and associated values. These methods will all be examples of supervised machine learning, because the (preprocessed) data itself will provide examples of the class values to be predicted as an outcome, while the other values might be used as features that might predict that outcome through some combination. SML methods generally assume (or work best when) features are independent (and thus cannot be redundant), values for features are always available, and the distribution of values

is uniformly balanced across their ranges, which is not always the case for medical data.

To address the complexities of real data, methods of feature selection and feature augmentation can be applied before the final SML models are created. SML models overall can be evaluated using either a fixed split between training and test sets; or by using a variant of cross-validation, which partitions the data over multiple evaluations (by random or stratified sampling technique), each of which uses a different partition as the training and test sets.

2.1.2.1. FEATURE SELECTION

Feature selection picks the best subset of features that maximize the performance of the model. Generally, this is done by evaluating different feature subsets- formed from leaving some out and testing them. This means we will know what the features to use, but not what values for features lead to each class in the decision problem. The aim of this procedure is to minimize the number of selected features to the optimal ones that explain the model the most. Among other benefits, feature selection removes redundant independent factors, while potentially preventing / minimizing 1.) errors; 2.) computational cost of measurement; 3.) underfitting; 4.) overfitting. The process also mitigates false or unreliable results. Feature selection adds value beyond model accuracy because it provides meaningful insight into the data, making the results of SML more transparent. Moreover, most modeling algorithms cannot efficiently handle high dimensional data, hence, it is reasonable to eliminate features that offer no value to the model [75]. In an imbalanced dataset, preceding data resampling with feature selection is suggested to be more effective, especially, if using SMOTE [76]. Four FSTs assessed in this study are: Recursive Feature Elimination, Recursive Feature Elimination with Cross-Validation, Pearson Correlation Coefficient and Variance Inflation Factor Analyses, and Genetic Algorithm.

2.1.2.1.1. RECURSIVE FEATURE ELIMINATION

RFE is an embedded FST that works with either deterministic (fixed) or stochastic (random) threshold. The threshold is the number of reduced features (n_features) to be used by the classifier. These cut off points serve as the stopping criteria for the iterations. The algorithm uses a classifier; hence, each iteration produces a model accuracy, and the index number of features is displayed alongside with the model accuracy. The algorithm is slow, but it has a track record of success in studies such as the identification of gene associated with myocardial infarction [77]. The authors explored RFE with SVM.

2.1.2.1.2. RECURSIVE FEATURE ELIMINATION WITH CROSS-VALIDATION

This is an ensemble feature selection approach. It aims is to identify and choose a unique set of optimal features obtained from the combined selections of multiple FSTs. Hence, the RFECV algorithm is a model based FST, thus operating through an estimator (classifier). Any type of cross-validation method can be implemented in RFECV, but the StratifiedKFold (SKFCV) is a robust cross-validation variant for imbalanced data through its sampling technique (stratified). The three methods involved in the FST are available through the Scikit Learn Application Programming Interface (API).

2.1.2.1.3. GENETIC ALGORITHM

Genetic Algorithm belongs to the wrapper category of feature selection approaches. It is an advanced biology-based (evolution) computational method that treats feature selection as a search problem, searching iteratively to find the overall optimal set of values for some parameter by generating new candidate combination (features) and evaluating them together, such that by the final stages of the search, only the combinations that lead to the highest quality results remain. This feature selector is widely used in computerized modeling tasks involving medical problems. The algorithm also has an impressive performance track record [78].

2.1.2.1.4. PEARSON CORRELATION COEFFICIENT AND VARIANCE INFLATION FACTOR ANALYSES

This method involves three sequential steps 1.) feature-feature multicollinearity check; 2.) feature-target correlation analysis, and 3.) Variance Inflation Factor (VIF) analysis for further assessed for multicollinearity assessment. Each step requires setting a threshold for correlation coefficient (r), and an R-Squared threshold for VIF. The output of one step is the input of the next step. The VIF method is an inverse of 1-(R- squared), and it is available through the Scikit Learn or Statsmodel library of Python programming language. The final output of VIF analysis becomes the selected features of the PCC-VIFA feature selection technique.

Mathematically, VIF is expressed as: *(1/(1-Ri^2))* (1)

,where Ri^2 is the amount of variance in y that is explained by x

2.1.2.2. DATA AUGMENTATION

The Synthetic Minority Oversampling Technique is a data resampling algorithm that is commonly used to resolve severe imbalances in the training set, such as when the number of positive examples is less than a quarter of the total, which would allow a training algorithm to achieve high measures of performance simply by defaulting to the majority class. There are different strategies utilized by the resampling method, the default being "auto", which ensures the number of the minority class and that of the majority class are the same. The K-Neighbors algorithm is the underpinning mechanism of the SMOTE algorithm (resampler). The default value for the k_neighbors parameter (k) in SMOTE is 5. The resampler can be implemented externally or internally through its Imblearn pipeline, and the algorithm can be optimized to enhance its effectiveness.

2.1.2.3. MODEL OPTIMIZATION METHOD

2.1.2.3.1. HYPERPARAMETER TUNING

Hyperparameter optimization is a process of finding the best combination of model parameters and their values towards minimizing errors and constructing an optimal model. Model optimization can be performed manually or automatically. The GridSearchCV (GSCV), from Scikit-Learn, is one of the algorithms that is commonly utilized to objectively tune (or regularize, in the case of logistic regression) the hyperparameters of a given classifier. The user provides a list of different hyperparameters and their corresponding values of a classifier to the "grid-param" parameter of the GSCV. The tuning algorithm then iterates over the list, and produce the best combination of parameters, based on a specified model performance scoring metric. The process is, however, time-consuming and the set of parameters (and their values) to explore are based on user's preference, leaving room for variation.

A key parameter optimization in the Tree-based algorithms (RF and DTree) is the "criterion", which determines the type of method the algorithms use for finding the informativeness of each feature in their decision making of class assignment. This concept is known as Information Gain (IG) analysis. RF and DTree use two approaches (Gini and Entropy) for calculating feature information gain, also known as feature importance. Entropy was selected by the GSCV method for both tree-based algorithms in this study.

2.1.2.3.2. ENTROPY

Entropy is a measure of the informativeness of a given feature to the determination of the class assignment by a Tree-based algorithm. Its score ranges between 0 and 1. The smaller the Entropy score of a feature, the optimal the split at such node; and a given node is said to be pure when its Entropy score = 0. When both classes in a binary classification obtain the same probability score in a feature, a maximum Entropy (1) score is obtained, and the node is said to be "impure", implying uncertainty. Classifiers are often optimized to maximize purity and reduce errors.

Formula for Entropy:

$$Entropy = -\sum_{j} p_j \log_2 p_j$$ *, where p represents probability* (2a)

For our binarized target, RF and DT would compute the Entropy score of each feature as follows:

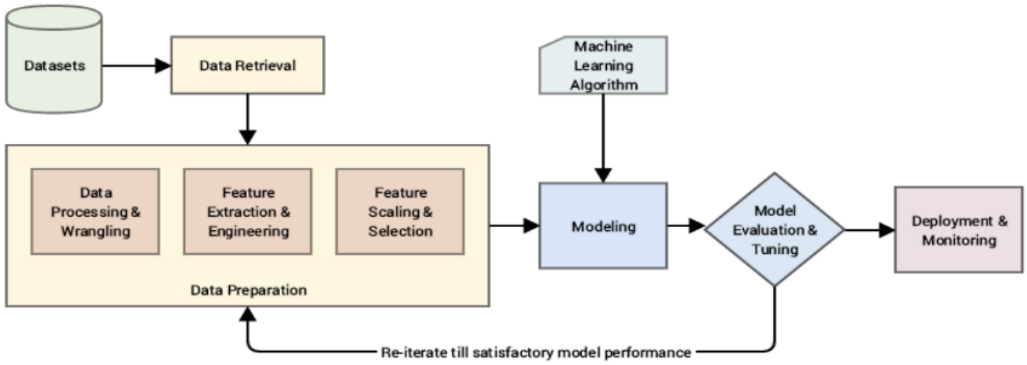
*Entropy=[(prob_Yes_GDHP)*log2(prob_Yes_GDHP)]-[(prob_No_GDHP)*log2(prob_No_GDHP)]* (2b)

2.1.2.4. MODEL EVALUATION METHOD

2.1.2.4.1. STRATIFIEDKFOLD CROSS-VALIDATION METHOD

The StratifiedKFold cross-validation model evaluation method is a type of cross-validation (CV) method that splits the dataset into k equal mutually exclusive stratified subsets, rather than performing a random split as K-Fold cross-validation does. K-1 folds are then utilized for training the model while each of the subsets is used exactly once for testing the predictive performance of the model. This process is repeated (iteration) k-times, and the model performance results per k-fold are then aggregated. The mean (average) value is the generalization result per model evaluation metric. StratifiedKFold CV is a well-known model evaluation approach when using an imbalanced data.

2.1.2.5. PIPELINE



Source: Dipanjan Sarkar, Raghav Bali, Tushar Sharma. Practical machine learning with Python (2018).

Figure 1: Standard supervised machine learning pipeline.

Within the scope of standard machine learning, a pipeline (Figure 1) represents a clear workflow (step-by-step) of a modeling task. Among other benefits, the use of a pipeline promotes model scalability and updatability.

2.1.2.6. OVERVIEW OF THE SUPERVISED MACHINE LEARNING ALGORITHMS EXPLORED

The predictive model of GDHP was established based on five standard SML classifiers and one deep learning algorithm. The support vector machine, logistic regression, random forest, decision tree and StackingClassifier were the standard SML algorithms tested, while the KerasClassifier was examined from the DL subset of SML. The six algorithms were employed to automatically differentiate whether a pregnant woman belongs to the positive class or the negative class

2.1.2.6.1. SUPPORT VECTOR MACHINE

Support vector machines are a distance-based method for training a classifier. The method sets a decision boundary between examples represented as vectors such that the distances between the examples (for either class) and the boundary are maximized. SVM can handle both multiple continuous and categorical variables. Studies show the algorithm often produces higher-performing models than other classification algorithms, including decision tree, and the conventional statistical methods [79]. This classifier uses regression to find and construct hyperplanes in a multidimensional space that best separate cases of different class labels, binary labels in this case. The farther away a point is from the separating line, the more confident one can be about the prediction for that point. In a study of concurrent Type 2 diabetes mellitus and hypertension in the general population, SVM was one of the two classifiers with outstanding performance [80].

2.1.2.6.2. LOGISTIC REGRESSION

Logistics Regression, also known as logit, is a statistics-based supervised machine learner that is used both in classification and regression problems. In binary classification, the algorithm predicts a probability for the positive class that can be mapped to a binary outcome by setting a threshold. The value of the threshold can be fixed (standard approach); determined experimentally or computed using various approaches. The standard approach is the default setting.

The LReg algorithm is widely employed for both classification and regression problems. In a binary classification problem, like the one in this study, LReg predicts the probability of a given instance (a data entry) belonging to a certain category or not. The Liblinear library is among other types of solver in LReg. The solver is geared towards binary classification, small to moderate sample size, and it supports L1 (Least Absolute Shrinkage and Selection Operator, also known as LASSO) and L2 (Ridge) regularization techniques. By default, Scikit Learn utilizes L2 as the regularization method in its logistic regression. LReg uses a non-linear activation function, known as the sigmoid function, to classify data. Many machine learning studies that model obstetric health problems such as GDM [81] and metabolic syndrome [19] successfully explored logistic regression.

2.1.2.6.3. RANDOM FOREST

The Random Forest classifier is a tree-based supervised ensemble machine learning algorithm that builds multiple decision trees from randomly explored and selected subsets of its training data, then combines their outputs. It then harnesses the power of the trees by aggregating their predictive votes to establish the final class of the test object. This characteristic of RF contributes to its advantages over a single decision tree. The RF

classifier is known for its high specificity and sensitivity; robustness to noise and outliers; and rare overfitting or underfitting; two problems that may arise from noisy data. RF is also computationally and time efficient (cost). Literature widely points to the optimality of combining many estimators as a classification algorithm [82], [75]. Among many clinical problems, RF was implemented for modeling in an acute kidney injury study [83]. Also, in a comparative study of diabetes where machine learning models were developed, both the LReg and RF classifiers performed similarly [84].

2.1.2.6.4. DECISION TREE

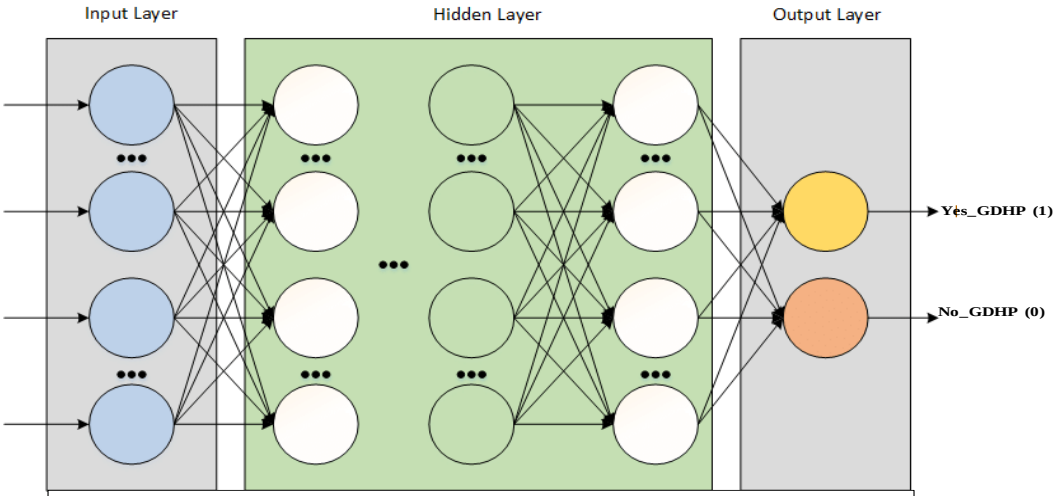
Another well-known Tree-based algorithm that is commonly used for supervised machine learning is Decision Tree. It estimates or constructs a single tree that can be represented in a flowchart-like tree structure. The topmost node in a decision tree is the root node, the internal node represents feature (attribute), the branch connotes a decision rule, and each leaf (terminal) node symbolizes the outcome. The classifier learns to partition the tree based on the attribute value in terms of how much information a feature can provide the sub-tree, thus calculating the Entropy (Section 2.1.2.2.3.2) or the Gini Impurity of such feature (IG analysis). It splits the tree using recursive approach (recursive partitioning). The flowchart-like structure assists in understanding the decision-making process of the algorithm as the attributes that weighed most heavily in the classification are transparent. Studies indicate one of the pitfalls of machine /deep learning model utilization in the clinical domain is the lack of human interpretability of some of the algorithms, hence, disease models are sometimes seeing as a black box [85]. In general, tree-based estimators ameliorate this barrier as they balance human readability and interpretability with efficiency, thus, adding value to the classification model. This characteristic bestows

a wide acceptance advantage on tree-based models among the medical experts [86]. The Decision Tree estimator has been employed in the health domain involving binary classification task to identify the relationships of several predictors to an outcome variable- metabolic syndrome [19], and GDM [87] for instance.

2.1.2.6.5. STACKINGCLASSIFIER

The StackingClassifier is a heterogeneous ensemble algorithm. The classifier combines multiple classification algorithms through a meta-classifier. The base estimators, also known as the first-level learners or weak learners, are trained on the entire training set. The algorithm is implemented using cross validation. The meta learner (second-level learner) gets fitted on the outputs of the base learners and the final prediction is then computed by the meta-estimator, hence. There are two ways of utilizing Stack– training the meta classifier on the predicted class labels or on the probabilities from the ensemble.

2.1.2.7. DEEP LEARNING ARCHITECTURE



Adapted from: Adhi Tama, Bayu & Rhee, Kyung Hyune. (2017). Attack Classification Analysis of IoT Network via Deep Learning Approach. Research Briefs on Information & Communication Technology Evolution (ReBICTE). 3. 10.22667/ReBiCTE.2017.11.15.015.

Figure 2: Deep neural networks architecture

A deep architecture/deep neural network (Figure 2) is a hierarchical, interconnected multiple layers of three major processing units- input, hidden (more than one hidden layer), and output. The concept simulates the human nervous system, thus having many nodes (neurons).

2.1.2.8. OVERVIEW OF DEEP LEARNING ALGORITHM

2.1.2.8.1. KERASCLASSIFIER

The KerasClassifier is a is Python compatible high-level deep neural networks API. It runs on top of TensorFlow, and its core data structures are layers of nodes, each of which has an activation function and parameters that are modified during training to create an optimal predictive model based on the training data. KerasClassifier is a Keras wrapper library from Scikit Learn that can be used for building deep network models for classification. The Keras-based classification algorithm follows the typical DL architecture; hence, it is a multi-layer (input, hidden, and output) artificial neural networks (ANN) variant. ANN, popularly known as "neural nets", is a biology-inspired data processing concept. There are various types of activation techniques (functions), but the rectified linear unit (ReLU) and sigmoid functions dominate in binary classifications. Their potential performance could be assessed over any number of epochs, as specified by a hyperparameter given by the researcher. ReLU and Sigmoid functions are linear and non-linear activation functions respectively. ANN algorithms have shown demonstrable success in the obstetric modeling, such as [88], [78].

$$ReLU\ function:\ f(x) = max\ (0, x) \quad (3)$$

Deep learning is a subdomain of supervised machine learning, and ANN modelling that has seen increasing use because it does not make assumptions about the independence or distribution of feature values and hence can do well even without feature selection or data augmentation. Deep

learning can also create decision models where relationships among features are non-linear, but using multiple layers of nodes that each use the training data to learn optimal weights for combining the inputs from other layers.

2.1.2.9. MODEL PERFORMANCE EVALUATION METRICS

Research supports using multiple measures to gauge and summarize model performance as no single measure captures all the attributes of a model [89]. The techniques to use for such assessment include model accuracy, F-score, precision, recall (sensitivity), and the area under the receiver operating characteristic curve.

Key:

- Positive (P): class instance is Positive
- Negative (N): class instance is Negative
- True Positive (TP): class instance is Positive, and the model predicted it as Positive.
- False Negative (FN): class instance is Positive, but the model predicted it as Negative.
- False Positive (FP): class instance is Negative, but the model predicted it as Positive.
- False Positive Rate (FPR)
- True Positive Rate (TPR)

2.1.2.9.1. PRECISION

Precision, also known as Positive Predictive Value (PPV), is the proportion of the correctly predicted instances as belonging to class **c** among all class instances of which the classifier claims that they belong to class **c**. So, this measure evaluates the fraction of correctly classified instances among the instances classified as positive. It is mathematically expressed as:

*Precision= TP/(TP+FP) *100* (4)

2.1.2.9.2. RECALL (SENSITIVITY)

Recall is the ratio (percentage) of the total number of class instances correctly classified as positive instances, divided by the total number of class instances correctly classified as positive, plus total number of class instances correctly classified as negative instances. High recall score indicates the classifier has correctly identified many True positive class instances and a small number of FN. Recall can be mathematically as:

*Recall = TP/(TP+FN) *100* (5)

2.1.2.9.3. MODEL ACCURACY

The accuracy of a given algorithm is the overall correctness of the model. It refers to the proportion of correctly predicted instances for each class to the total number of sample cases. Therefore, the accuracy, ***A***, of algorithm, ***m***, can be mathematically expressed as:

*Am=(s/N)*100* (6a)

*, where **s** is the sum of correct predictions, and **N** is the total number of predictions made.*

Or, [(FP +FN)/(TP+TN+FP+FN)] (6b)

It is cautionary to indicate that model accuracy could be a misleading model evaluation metric, especially with an imbalanced data. So, the final decision about model performance is not commonly based on this measurement.

2.1.2.9.4. F1

The F1, also known as F-Measure, represents both precision and recall. It is the harmonic mean of precision and recall. It uses the harmonic mean in place of the arithmetic mean to regularize the extreme values more. F1 is mathematically expressed as:

$$F = (2*(Precision * Recall) /(Precision + Recall)) \quad (7)$$

2.1.2.9.5. AREA UNDER THE RECEIVER OPERATOR CHARACTERISTIC CURVE

The area under the receiver operator characteristic curve of a model is an equivalence of the c-statistic. This metric is a probability curve of the model ability to separate the positive class from negative class. The measure is commonly being referred to as, AUC or ROC Curve. It is used to display the performance of a binary classification algorithm, but it could be modified for a multiclass as well. The closer the AUC score to 1 (maximum obtainable score), the better; and that shows the model is a well-performing one with a high chance (probability value) of good differentiation capability between positive and negative classes. A poor model would have its AUC near 0; implying the algorithm has a worst measure of separability (no class is selected). AUC score of 0.5 denotes the model is possibly merely interchanging the result by predicting 0s as 1s and vice vasa. ROC Curve is normally plotted with TPR against the FPR where TPR goes to the y-axis and FPR is on the x-axis.

2.2. RELATED STUDIES

2.2.1. RELATED MACHINE LEARNING BASED ANALYSES TO PREDICT GDHP

One of the few machine-learning based studies of GDHP has been that of Du et al (2020) [90], where deep learning was the methodology of analysis. The authors conducted image analysis of

unstructured data (radiology dataset; n=548) of fetal lung to investigate GDM and preeclampsia individually and concurrently with the goal of predicting neonatal respiratory morbidity (NRM). The focus of the study clearly differs from that of the present study as it was more fetal-oriented than maternal, the dataset analyzed was unstructured data, and the sample size was small. However, the current study centers on maternal risk identification for GDHP, and the commonly obtained prenatal data was examined.

In a nationwide prospective study [82] carried out in Indonesia in the year 2020, the authors adopted artificial intelligence (AI) techniques for a preliminary prediction of preeclampsia. The study analyzed the data of the preeclamptic/eclamptic (n = 3318) vs pregnant women with normal blood pressure (n = 19,883) with singleton pregnancy. There were 95 features in the dataset, ranging from demographic data to past medical histories (from 24 months prior to the event, to delivery as the event). Feature selection found only 17 predictors to be the most influencing factors as identified by random forest. We plan to test this algorithm out in feature selection processes of the GDHP modeling as well. SVM, LReg, DTree, RF, artificial neural networks and an ensemble learner (a combination of all other mentioned algorithms) were the six classification algorithms explored by the authors. AUC was employed to compare the models. Finding from the study indicates pre-conceptual health around one year ("9-12 months to the event") is crucial to perinatal outcomes. Using precision, sensitivity, and specificity in the validation sets, the model built in Sufriyana et al.'s work [82] outperformed the existing ones. Although this obstetric modeling yielded a very informative result about preeclampsia prevention, excluding women with other variants of HDP and the gestational diabetic hypertensive (or vice versa) women is one of its limitations. This is especially the case;

considering the assertion that the factors responsible for HDP variants are very similar, and that the conditions carry an equal level of risk, which includes cardiovascular disease [4].

Additionally, SML methods were applied to a genome-wide association study (GWAS) conducted by Kong and Choe (2019), where a female-specific metabolic syndrome predictive model was constructed (n=3,968) [91]. The GWAS research was devoted to finding fundamental genetic concepts in relation to MetSyn of the female gender. Model evaluation depicts the AUC of the ROC curve for female is significantly higher (AUC = 0.85) than that of male (AUC = 0.57). The results are consistent with literature as they imply that women are genetically more susceptible to MetSyn than men [15], [92]. The study, however, was data-restrictive because genetic risk factors were solely explored in the prediction of metabolic syndrome affecting women, while studies show disparate non-causal factors are responsible for HDP. Also, even though the study focused on women and MetSyn, pregnancy-related factors were not examined.

2.2.2. RELATED NON-MACHINE LEARNING BASED ANALYSES TO PREDICT GDHP

Earlier data-driven work on GDM and HDP co-occurrence did not use machine learning. Ling et al. (2018) [93] carried out a prospective study using conventional statistical methods to ascertain the individual and synergistic effect of GDM and HDP on postpartum cardio-metabolic risk. Of the 400 pregnant women who were recruited at their early pregnancy period (5-8 weeks), 276 of them eventually participated in the 5-year follow-up study. The authors described MetSyn as an abnormal glucose metabolism and hypertension. The results of the study show HDP and GDM individually and collectively impair postpartum cardiometabolic health. The study reported a Relative Risk of 2.6 (1.7-3.9) and 2.7 (1.6-4.9) for each case (individual condition and jointly)

respectively. Although, there is a thin line of difference between these relative ratios; with both having approximately threefold risk ratio, their slight difference cannot be overlooked. Meanwhile, due to the inherent drawbacks of the study design (longitudinal study), the result is not generalizable. For instance, there is a 5-year gap between the time when the study commenced and when it ended. Apart from the study recording nearly 50% (124 gravidae) response loss, much could have changed in the participants' health status (and /risk factors) after the study started (during pregnancy), and after delivering their babies. For instance, some could have embarked on lifestyle modification, as various studies indicate these changes are necessary to prevent or minimize maternal cardiometabolic risk. The final sample size (276) is also too small for the finding to be applicable to the pregnant population.

Another non-SML study within the paradigm of concurrent GDM and HDP was conducted in 2018 by Cao et al. [94]. The co-existence of gestational diabetes mellitus and preeclampsia using the laboratory data (C-reactive protein (CRP) and interleukin-17/IL-35) was used, along with the BMI of four groups of pregnant women (139). The groups were assembled according to attributes, including preeclampsia, normal blood pressure, GDM and co-existing GDM and preeclampsia. Multiple conventional low-level data analysis methods were applied, including statistical significance of difference and Analysis of Variance with Post-hoc Turkey's test to assess the difference of means of the four groups. Pearson correlation coefficient was also applied to examine the association between predictors. Its result shows there is a positive correlation between the following variables diastolic blood pressure with interleukin (IL)-17 levels; BMI and triglyceride; and, between IL-17 levels with BMI and proteinuria in the group with comorbidity of GDM and preeclampsia. Results from the analysis demonstrate maternal hyperlipidemia (lipids), hyperglycemia, high BMI, high CRP levels and imbalanced interleukin-

17/IL-35 may lead to the comorbidity of GDM and PE. This insight notwithstanding, the sample size is too small, and its scope is limited. A myriad of factors determining the co-existence of GDM and preeclampsia were left out by analyzing only maternal laboratory and BMI data. This is especially so because as Gaillard et al.'s work [73] reveals, laboratory tests may fail in certain situations; and its ability may not extend to certain risk factors of these diseases. However, the finding from Cao et al.'s work is consistent with the other research indicating GDM and HDP interrelate. It also shows the impact of GDHP extends beyond the pregnancy period.

2.2.3. SUMMARY OF RELATED DATA-DRIVEN STUDIES ON GDHP

As the discussed past machine learning research have shown, supervised machine learning clearly holds promise for identifying pregnant women who are at risk of developing GDHP. Generally, machine learning research centering on the comorbidity of GDM and HDP is rare. Therefore, the current study adds to the body of knowledge, and its superiority is in in many folds 1.) it is more comprehensive in terms of the disease component as it considers the six HDP fields and one GDM field that are available in WiseSubset to create the comorbidity field; 2.) it is also wide-ranging in methods because it simultaneously assessed the utility of five standard SML and a deep learning algorithm on routine prenatal care attributes. To the best of our knowledge, there is no published model that combines such methods to study GDHP; and 3.) this is also the first research to classify the gravid population as having GDHP or not, using a PeriData set or otherwise, and regardless of the methodology. These observations point to the uniqueness of our research in the domains of machine learning and obstetrics.

3. METHOD

3.1. INSTITUTIONAL REVIEW BOARD

We obtained an Institutional Review Board (IRB) approval (IRB No.21.250) from the University of Wisconsin, Milwaukee before embarking on this research. We also had Data Use Agreements with the clinical site. To be Health Insurance Portability and Accountability Act compliant, we obtained a waiver for the two PHI fields in our research protocol. For the purpose of this research, we named the sample as, “WiseSample”; and the extract of WiseSample was called, “WiseSubset”. Some of the fields in the sample were smoking, prepregnancy weight, fetal gender, mother’s blood group, mother’s calculated age, and mother’s reported age.

3.2. DATA PREPROCESSING

3.2.1. EXPLORATORY, DESCRIPTIVE DATA ANALYSIS AND DATA EXTRACTION

A custom program was written in the Python programming language to analyze the WiseSample dataset for the binary classification of GDHP. We started the exploration of the WiseSample dataset with 9962 data entries, each with 79 fields. We first renamed all the fields, and certain value names that were not compatible with some of the data processing methods utilized in this study were converted to numeric values. Exploratory analysis showed the sample was noisy, and placeholder values were identified. We then replaced the placeholders with the token “Nan” for missing value (MV)s, and filtered the DataFrame by excluding from the sample all instances that lack an actual value for the following seven outcome fields used to determine the target (comorbidity) field: Preeclampsia (2 fields), with or without severe features; gestational hypertension (2 fields), from the maternal or from the child record; eclampsia (1 field); HELLP

syndrome (1 field); and gestational diabetes (1 field). This resulted in WiseSample having 4794 instances and 79 fields.

3.2.2. MISSING DATA ANALYSIS AND FEATURE ENGINEERING

There were fields with missing values or low frequency count values in WiseSample. We merged some missing value fields to create a composite field appropriately. Such fields include 1.) the 13 substance-abuse-related fields in the sample (formed Combo_Subst_Use field); smoking and secondhand smoking (created smoking_FirstSec field). In some fields with low frequent count values, such as mother's primary race, we merged such values to create a combined value in their respective fields. This data manipulation increased the number of fields to 82 before we dropped the original fields that were combined. We set and utilized a data imputation threshold of a MV $\leq$ 30.00%, then applied the SimpleImputer method of Scikit Learn to address the numeric and categorical missing data. We then dropped fields not meeting the imputation cutoff point; the redundant fields; past obstetric history fields that were suggestive of previous GDM/HDP (a delivery history of macrosomia for instance, indicates previous GDM); and irrelevant fields (e.g. hysterotomy). In this phase, we were left with 4794 instances and 34 fields in the sample.

3.2.3. DATA EXTRACTION

We applied two study participation eligibility criteria: no prepregnancy diabetes, and no chronic hypertension, and 4624 instances fulfilled the requirement, hence, WiseSample had 4624 instances and 34 fields. Next, we engineered a binary comorbidity field, GDHP" (Yes [1]/ No[0]) by merging the seven GDHP-related outcome fields, (Section 3.3.1), thus extracting the WiseSubset dataset (4624 instances and 36 fields) from WiseSample. We then dropped the

merged primary fields; the participation criteria; and the merged fields that persistently have low frequent counts. The WiseSubset dataset then had 4624 instances and 24 fields.

3.2.4. DATA TRANSFORMATION

To normalize the numerical data, we utilized the MinMaxScaler method of Scikit Learn. We transformed the continuous data in multiple steps into categorical data to engineer new fields from their respective values. Mother's reported age, husband's reported age, prepregnancy weight, and pre-pregnancy BMI were among the fields that benefitted from this operation. Ordinal data such as mother's educational status were transformed numerically to preserve and reflect their natural order, while the nominal data was transformed into numerical data using the LabelEncoder method of Scikit Learn as well.

3.2.5. FEATURE ENGINEERING

We carried out multiple feature engineering operations to create new fields from existing fields where necessary, to reduce the sparsity among value types for some features of the WiseSubset dataset and to add features that are implicit in the data, but not represented directly. Among such operations was the creation of a "Season" field from the month part of the "date of first prenatal care visit" field, and its four weather (United States) values were later turned into four new binary fields. Also, the "Year" in the "date of first prenatal care visit" field was utilized with the "Year of Last birth" field to engineer an interpregnancy interval field. The field of mother's height in feet and the inches part of the height value, which was a separate field were merged to create a single field (feet.inches). We created two categorical fields from the new field. Additionally, we followed the adult BMI classification of the CDC [95] to categorize the prepregnancy BMI field, then, engineered five new fields from its five values. Mother's reported

age, and husband's reported age were also decomposed into three age ranges respectively. Another field that benefitted from this phase was prepregnancy weight, from which we created three new fields (small, medium, large). We similarly transformed the gestational week at first prenatal care visit field into three fields. In total, 24 new fields were derived from the existing fields in this phase, we call them, "TheNew24". We dropped all primary/intermediate fields that were utilized during feature engineering, then, WiseSubset finally consisted of 4624 instances with 38 features and 1 target field.

3.3. FEATURE SELECTION

We assessed the utility of four different feature selection techniques to select the optimal feature subsets from the 38 features while the dataset was still imbalanced. The Scikit Learn pipeline was used to implement the embedded FSTs: Recursive Feature Elimination, Recursive Feature Elimination with Cross Validation, and the wrapper method (Genetic Algorithm). Pearson Correlation Coefficient and Variance Inflation Factor Analysis was also examined.

3.3.1. PEARSON CORRELATION COEFFICIENT AND VARIANCE INFLATION FACTOR ANALYSES

We started PCC-VIFA by obtaining a correlation matrix of the entire feature set (38), then, we divided the features into three themes: Obstetric, Maternal-paternal profile, and Environmental, and we obtained their correlation matrices. We then set three different thresholds for each of the three sequential steps in this FST: feature-feature multicollinearity check, feature-target correlation analysis, and an advanced multicollinearity analysis with VIF with the Statsmodel Python library. For feature-feature pairwise correlation, we utilized a correlation coefficient (r) threshold of abs($r < 0.75$), thus giving considering both the positively and negatively correlated features. We similarly applied a threshold of abs($r < 0.00$) in the feature-target correlation. The

intermediate result was an input data for VIF analysis, with a threshold: R Squared= 2.50. Features not meeting the cutoff point were automatically dropped. The 19 features that were finally selected by PCC-VIFA were named, “19PCC-VIFA”, and we later subjected them to information gain analysis with random forest. The selected features were the input data for, “Model Set 4”.

3.3.2. RECURSIVE FEATURE ELIMINATION

To implement RFE, we randomly split the data into 8:2 ratio of train and test sets. Then, we utilized the GridSearchCV algorithm to find the best parameters and their values for the GradientBoosting **(**XGB**)** classifier, which we used as an estimator inside the pipeline of the RFE algorithm, hence, RFE-XGB. We then fitted the pipeline on the training set and iterated over the range of the indices of the 38 features, starting from index 1. Model performance was generated per iteration as the model made predictions on the test set. No GDHP model was constructed through this FST.

3.3.3. GENETIC ALGORITHM

We first randomly split the data into 8:2 ratio of validation set, and TrainAndTest sets, the latter was then split randomly as train-test sets. The genetic algorithm framework employed was the Distributed Evolutionary Algorithms in Python (DEAP). We set the initial population size to 0, initialized the weight with 1 and 0 for bias, and we set population size to 100 per generation of 10. We implemented the tournament selection method for choosing individuals from a population. The crossover probability was also set to 0.5, and 0.2 for the mutation probability. A binary vector of 0 and 1 was created for all the features, where 1 implies the corresponding feature would be selected by the estimator, and 0 otherwise. We wrapped genetic algorithm on logistic regression, hence,

Gen-LReg, to build a model per iteration. A test accuracy score (fitness score) was obtained per iteration through a fitness function that evaluates each "individual" (combination of features). The Genetic Algorithm stopped when the population converged at an optimal solution, and that was an individual containing 22 features. We named the predictors, "22Gen", and obtained their importance scores with random forest. A barh chart (a horizontal bar chart) was also generated to visualize their ranks. We called the model set that utilized the "22Gen" as input data, "Model Set 5".

3.3.4. RECURSIVE FEATURE ELIMINATION WITH CROSS-VALIDATION

We utilized 10-fold StratifiedKFold cross-validation to satisfy the cross-validation parameter of the RFECV, while we passed random forest as its estimator, thereby creating an ensemble feature selection. This method selected 9 features that we called, "9RFECV-RF". We obtained a Line chart for visualizing the CV scores and their corresponding numbers of selected features. A barh chart was also plotted for feature importance with RF. We utilized the 9RFECV-RF for constructing, "Model Set 6".

3.4. DATA AUGMENTATION

We utilized SMOTE to resolve the class imbalance noted in the WiseSubset data exploration phase. For the four (LReg, DTree, RF and SVM) of the five standard classifiers, we implemented the resampling algorithm (SMOTE) on the "train" part of each of the stratified fold, thus preserving the gold standard ("test" part of each fold). This was done through the Imblearn pipeline. We, however, transformed the dataset externally with SMOTE when using Stack and Keras since they have no pipelines. In both SMOTE implementation types, the resampler was utilized both in its default and in an optimized mode. For clarity, we refer to the augmented

training data per mode as, “Augmented Dataset1” and “Augmented Dataset2” respectively. To optimize SMOTE, we tuned its k_neighbors (k) parameter by iterating over a range of k values (1-10) to select the best value of k in SMOTE per classifier. We set the strategy parameter to “auto” and examined the utility of the data augmentation algorithm through the models created with each augmented dataset. While Augmented Dataset1 was used to construct the models that we call, “Model Set 2”, its Augmented Dataset2 counterpart was utilized for creating Model Sets 3, 4, 5, and 6. Meanwhile, the Keras model was only included in Model Sets 2 and 3 with SMOTE because FST outputs were the data input in the remaining four model sets, and feature selection is not a separate process in DL.

3.5. CONFIGURING AND TESTING THE SUPERVISED MACHINE LEARNING MODELS

We examined various models created using the following six types of classifiers: Logistic regression, random forest, decision tree, support vector machine, StackingClassifier, including a DL classifier known as KerasClassifier. All the algorithms, except the latter, were utilized in both their default and optimized modes. For each classifier (except for Stack), we utilized an hyperparameter tuning algorithm, known as GridSearchCV, to objectively search and find the best hyperparameters and their values from a list of parameters and some corresponding values that produce a model with good performance, thus optimizing the classifiers. We then examined the configuration quality of the classification algorithms, using standard performance measures. We developed a set of base models (Model Set 1) while the LReg, SVM, RF and DTree were in their default states through their respective full Scikit Learn pipelines. Model Set 2 was similarly built with the listed classifiers still in their default states, but Imblearn pipeline was utilized, and SMOTE (default) was added. We constructed four different model sets (Model Sets 3-6) with the

optimized classifiers (and an optimized SMOTE), all implemented through the Imblearn pipelines of each of the classifiers. For the StackingClassifier, we set RF, DTree, and SVM as the base learners, while LReg was utilized as the final estimator (meta learner). This configuration was implemented and tested in 6 scenarios to build Model Set 1-6. The deep learning algorithm was developed with KerasClassifier. The model was configured as follows: 4 layers of 1 input layer (38 input data), 2 hidden layers, and 1 output layer. The layers had 16,16,16 and 1 node respectively: with a corresponding ReLU activation function, except for the last layer where Sigmoid was utilized as the activation function; an optimizer (Adam); a loss function (binary crossentropy); and a scorer (accuracy).

For building the Keras baseline model, we used 400 epochs and 2000 batchsize, and the imbalanced data was analyzed with the 38 features. We then tested the configuration with one case, the baseline (Model Set 1). Then, we included hyperparameter tuning in the DL classifier configuration of subsequent modeling by setting a range of epochs between 2 and 10, and batchsize to 100 and 1000, and incorporated GSCV to automatically test and update the respective values during learning. We tested the tuned configuration with Model Sets 2-3.

3.6. MODEL TRANSPARENCY WITH DECISION TREE

We constructed 33 decision tree models in this analysis. To promote transparency (Section 2.1.2.3.1), a single tree structure was obtained as a representation for visualization from the model set where the DTree algorithm recorded the best performance (Model Set 4), based on recall and AUC. This represents/ portrays the concept of model transparency that is common to medical modeling, and Entropy (Section 2.1.2.3.2) was utilized in the DTree algorithm.

3.7. SUMMARY OF EXPERIMENTS

All the six classifiers in this analysis utilized the full set of features in 18 scenarios (Model Sets 1-3), while we additionally tested each of the five standard SML algorithms using each of the four derived feature sets from the FSTs in 15 scenarios (Model Sets 4-6). Therefore, a total number of 33 predictive modeling scenarios were created in this binary classification of the WiseSubset dataset for GDHP. The Sklearn pipeline was used in 4 (LReg, RF, DTree, and SVM in Model Set 1) of the 33 tests, while the Imblearn pipeline was utilized 20 times. Testing Stack (6 cases) and Keras (3 cases) with no pipeline created the remaining 9 scenarios. The SKFCV cross-validation method was applied to evaluate every model constructed. To measure and summarize model performance, we utilized accuracy, F1, precision, recall (sensitivity), and AUC; ROC plots were also generated. All the models in this research were systematically designed and grouped by input data.

4. RESULTS

4.1. DATASET

The exploratory analysis of WiseSample showed the dataset contained actual, missing, and placeholder values. Missing data analysis also revealed 31.09% and 31.87% of the dataset had a missing value (before and after replacing the placeholder values with a token “Nan” for missing data). Part of the results (Figure 3) of the descriptive analysis of the WiseSample is the imbalanced distribution of the seven outcome fields that were utilized in creating the GDHP. This resulted in the class imbalance of the WiseSubset against the positive class as shown in the GDHP chart inside Figure 3. The imbalance is approximately ratio 1.1:100 for the positive class and the negative class respectively. Table 1 summarizes the descriptive results of the WiseSubset, and the new 24 fields created during feature engineering are displayed in Table 2. Also, a full list of the starting fields is available in Appendix A.

Table 1: Summary of descriptive analysis of WiseSubset

Summary Table: WiseSubset Preprocessing Output	
Descriptive Statistics	**Results**
Oldest year of first prenatal visit	2015
Latest year of first prenatal visit	2018
Multiparous	3381
Nulliparous	1243
Delivery of male infants	2404
Delivery of female infants	2220
Youngest mother (age in years)	15
Oldest mother (age in years)	44
Youngest husband (age in years)	20
Oldest husband (age in years)	50
Commonest maternal height (ft)	5
Positive examples	49
Negative examples	4575
New fields engineered from decomposed fields	24
Drop decomposed fields, then, WiseSubset finally contains:	4624 instances, 38 features, and 1 target field

Table 2: List of the new 24 features created in WiseSubset

Serial Number	24 Newly Created Fields in WiseSubset
1	Spring
2	Summer
3	Fall
4	Winter
5	Very_Young_AdultM
6	Young_AdultM
7	Middle_AgedM
8	First_Trim
9	Second_Trim
10	Third_Trim
11	Healthy_Weight_bmi
12	Overweight_bmi
13	Small
14	Medium
15	Large
16	Young_AdultH
17	Middle_AgedH
18	Older_AdultH
19	Average_height
20	Tall
21	Obese
22	Class_3_Obese
23	UnderWeight_bmi
24	InterPreg_Interval

4.2. FEATURE SELECTION

Using the RFE-XGB feature selection method, the test accuracy was 98.56% from index 1 to 38, hence, this method selected no feature and, thus, no model set. The three steps in the PCC-VIFA approach eliminated 6, 1 and 12 features respectively. The thresholds were abs($r<0.75$), abs($r<0.00$), and R Squared=2.50 accordingly; and PCC-VIFA eventually selected 19 features. The correlation matrices obtained are in Appendix C-F. The Gen-LReg method selected 22 features as the best explanatory factors for GDHP. The RFECV-RF strategy identified 9 best features, and a Line chart displaying the cross-validation scores per number of features selected is available in Figure 4. Also, a barh chart (horizontal bar chart) is in Figure 5 to visualize the importance (normalized feature ranking) of the “9RFECV-RF” selections. The results of the feature importance analysis conducted with random forest on each feature representative set

selected by the three FSTs that selected a feature subset is in a combined Table (Table 3). There is also a Table (Table 4) displaying the entire 38 features, and the selected features per feature selection method, together with their non-normalized ranks. However, Appendix B is a unified view of any information pertaining to all the features and their selections.

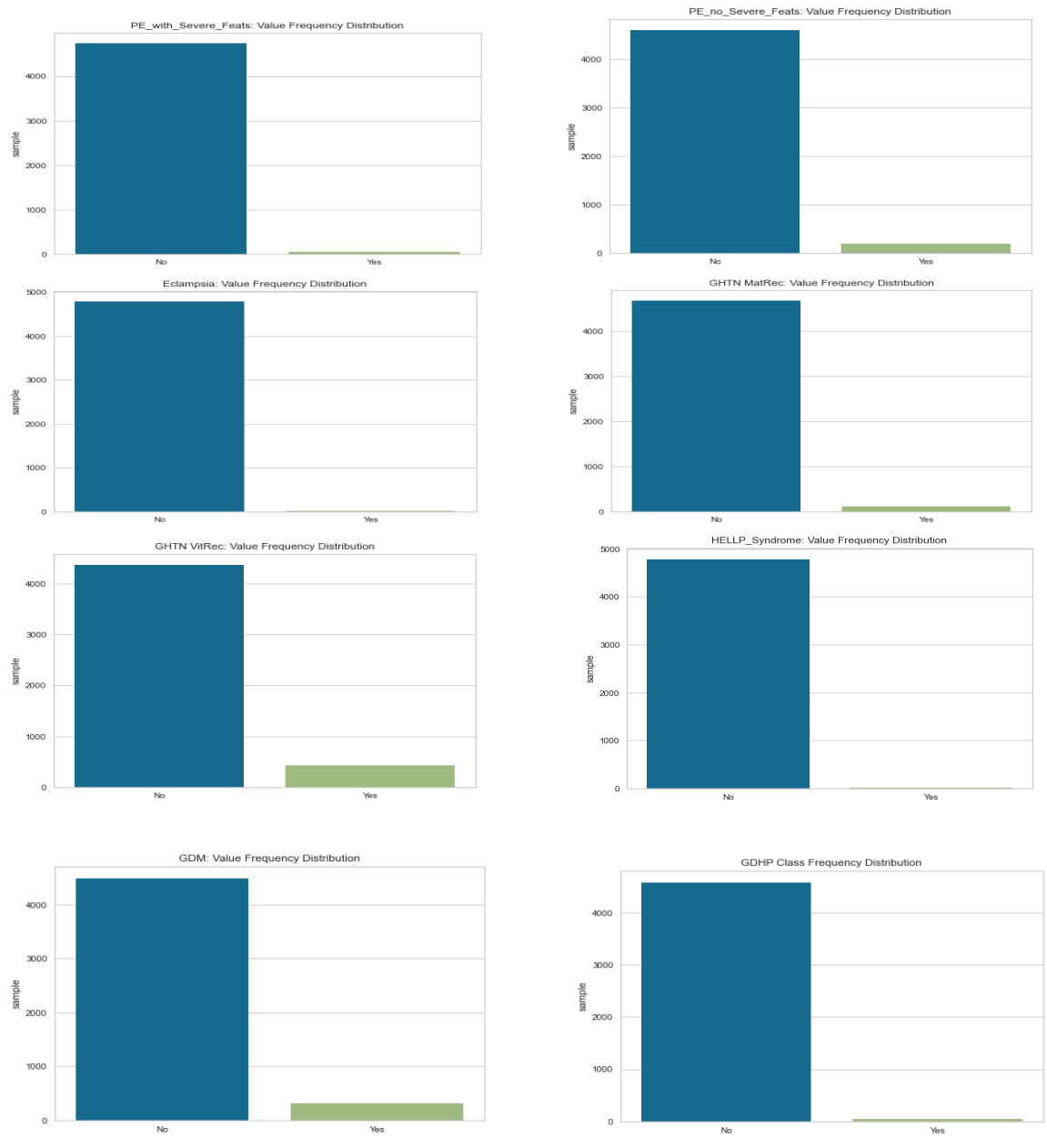


Figure 3: Multiple charts showing class imbalance of the GDHP field and its origins.

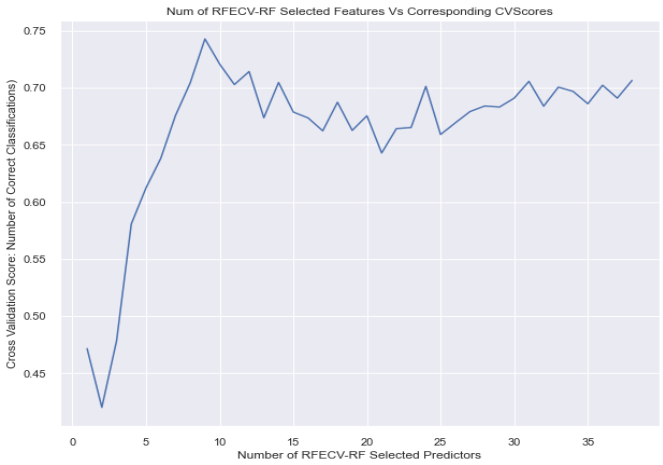


Figure 4: Line chart of the RFECV-RF feature selection process

Table 3: Normalized feature ranking (importance) of the selected features per FST

Serial_No	9RFECV-RF: Feature Name	IG_Score
0	InterPreg_Interval	0.1977
1	Mother_Edu_Status	0.1421
2	NumOf_Preg_PlusCurrent	0.136
3	Mother_Blood_Group	0.1077
4	Healthy_Weight_bmi	0.1068
5	Husb_Edu_Status	0.0888
6	Medium	0.0828
7	Middle_AgedM	0.0728
8	Husb_Occup_1yr_Ago	0.0652

Serial_No	19VIFA: Feature_Name	IG_Score
0	Mother_Blood_Group	0.1083
1	Class_3_Obese	0.0922
2	Pry_Payer_PrenatalCare	0.0852
3	Middle_AgedM	0.0835
4	Very_Young_AdultM	0.0803
5	Obese	0.0734
6	Average_height	0.0697
7	Infant_Gender	0.0622
8	Overweight_bmi	0.0567
9	Fall	0.0552
10	Smoking_FirstSec	0.0518
11	Summer	0.0477
12	Young_AdultH	0.0433
13	Winter	0.0413
14	Older_AdultH	0.0219
15	UnderWeight_bmi	0.0137
16	Third_Trim	0.0104
17	Combo_Subst_Use	0.0032
18	Small	0

Serial_No	22Gen: Feature_Name	IG_Score
0	Mother_Edu_Status	0.1203
1	Healthy_Weight_bmi	0.1097
2	Middle_AgedM	0.0865
3	Husb_Edu_Status	0.0846
4	Infant_Gender	0.0655
5	Husb_Occup_1yr_Ago	0.063
6	Large	0.0596
7	Class_3_Obese	0.0503
8	Smoking_FirstSec	0.0484
9	Young_AdultM	0.043
10	Pry_Payer_PrenatalCare	0.0421
11	Medium	0.0405
12	Winter	0.0318
13	Fall	0.0312
14	Second_Trim	0.0292
15	Older_AdultH	0.0256
16	Mother_RH_Status	0.025
17	First_Trim	0.0217
18	UnderWeight_bmi	0.0116
19	Third_Trim	0.0083
20	Combo_Subst_Use	0.0022
21	Small	0

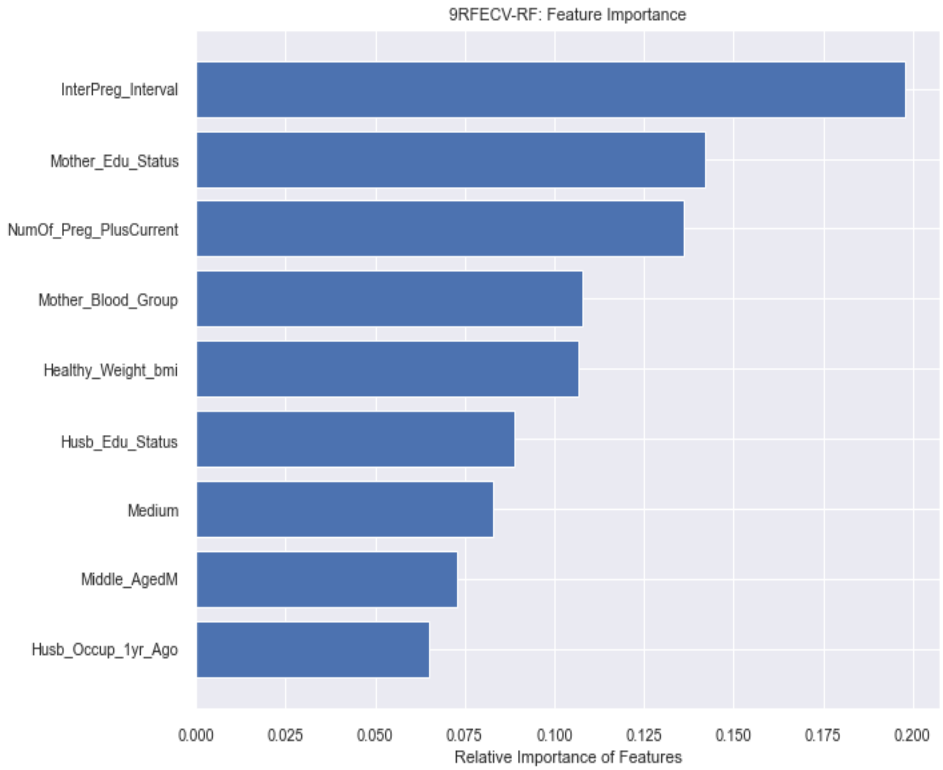


Figure 5: Barh chart showing the importance of the features selected by RFECV-RF

Table 4: All features, their selected subsets per FST, and their non-normalized ranks

Summary Table: Output of Feature Selection and Their Ranks (Non-normalized Version)

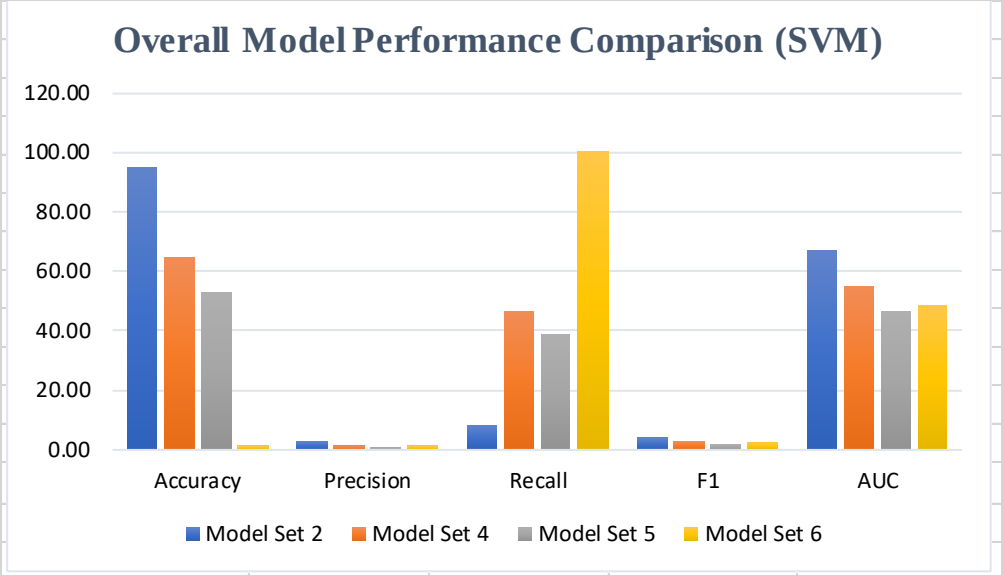
Key: 0= Unselected; then, ranking starts at 1, the smaller the number, the higher the rank out of 38
Green= Optimal FST, Pink= 6 of the 9 optimal features; also in the top 6 of ≥ 1 other FST output

Serial Num	Final Feature Set (38)	RFECV-RF	Gen-LReg	PCC-VIFA	RFE-XGB
1	Infant_Gender	0	5	8	0
2	Mother_Married	0	10	0	0
3	Mother_Pry_Race	0	0	0	0
4	Mother_Edu_Status	2	1	0	0
5	NumOf_Preg_PlusCurrent	3	0	0	0
6	Parity	0	0	0	0
7	Pry_Payer_PrenatalCare	0	11	3	0
8	Mother_Ever_Married	0	0	0	0
9	Husb_Edu_Status	6	4	0	0
10	Husb_Occup_1yr_Ago	9	6	0	0
11	Mother_Blood_Group	4	0	1	0
12	Mother_RH_Status	0	17	0	0
13	Combo_Subst_Use	0	21	18	0
14	Smoking_FirstSec	0	9	11	0
15	InterPreg_Interval	1	0	0	0
16	Spring	0	0	0	0
17	Summer	0	0	12	0
18	Fall	0	14	10	0
19	Winter	0	13	14	0
20	Very_Young_AdultM	0	0	5	0
21	Young_AdultM	0	10	0	0
22	Middle_AgedM	8	3	4	0
23	First_Trim	0	18	0	0
24	Second_Trim	0	15	0	0
25	Third_Trim	0	20	17	0
26	Healthy_Weight_bmi	5	2	0	0
27	Overweight_bmi	0	0	9	0
28	Small	0	22	19	0
29	Medium	7	12	0	0
30	Large	0	7	0	0
31	Young_AdultH	0	0	13	0
32	Middle_AgedH	0	0	0	0
33	Older_AdultH	0	16	15	0
34	Average_height	0	0	7	0
35	Tall	0	0	0	0
36	Obese	0	0	6	0
37	Class_3_Obese	0	8	2	0
38	UnderWeight_bmi	0	19	16	0

4.3. OVERALL MODEL COMPARISON

We used the following six SML algorithms; LReg, SVM, RF, DTree, Stack, and Keras for constructing predictive multiple binary classification models for GDHP. The results of the overall model comparison (Figure 6) show SVM recorded the best mean recall (100.00%) in Model Set 6. The optimal features were identified by the RFECV-RF feature selection method The ROC plot per model in Model Set 6 is available in Appendix G-K.

Figure 6: Results of overall model comparison



Metric(Mean %)	Model Set 2	Model Set 4	Model Set 5	Model Set 6
Accuracy	94.81	64.77	52.68	1.19
Precision	2.61	1.43	0.86	1.06
Recall	8.00	46.50	39.00	100.00
F1	3.87	2.78	1.69	2.10
AUC	67.03	54.98	46.20	48.62

4.4. MODEL PERFORMANCE COMPARISON BETWEEN ALL MODEL GROUPS

The results (Table 5) indicate Model Set 6 produced the largest mean recall score (100.00%) through SVM, using the 9 features selected by the RFECV-RF feature selection method. The highest mean AUC score recorded in this study was from Model Set 1, where LReg recorded 76.67% by utilizing the entire 38. Model Set 2 and 3 recorded the highest mean precision (5.00%) through LReg as well. In both model sets, the 38 features were utilized with the Augmented Dataset1 and Augmented Dataset2 respectively. For F1, its highest mean score was from DTree Model Set 2 (4.93%), utilizing the Augmented Dataset1 and the 38 features.

Table 5: Results of comparison between all the classifiers assessed

Comparison Between The Six SML Algorithms Explored in Predicting GDHP						
Classifier	**Model Group**	**Accuracy**	**Precision**	**Recall**	**F1**	**AUC**
SVM						
	Model Set 1	98.94	0.00	0.00	0.00	49.69
	Model Set 2	94.81	2.61	8.00	3.87	67.03
	Model Set 3	55.93	0.58	25.50	1.13	36.57
	Model Set 4	64.77	1.43	46.50	2.78	54.98
	Model Set 5	52.68	0.86	39.00	1.69	46.20
	Model Set 6	1.19	1.06	100.00	2.10	48.62
LReg						
	Model Set 1	98.94	0.00	0.00	0.00	76.67
	Model Set 2	98.57	5.00	2.00	2.86	65.14
	Model Set 3	98.79	5.00	2.00	2.86	69.67
	Model Set 4	80.02	1.53	26.50	2.78	54.98
	Model Set 5	85.99	2.05	22.00	3.68	69.73
	Model Set 6	73.47	2.31	57.00	4.44	48.19
RF						
	Model Set 1	98.94	0.00	0.00	0.00	60.67
	Model Set 2	98.92	0.00	0.00	0.00	64.71
	Model Set 3	98.70	0.00	0.00	0.00	72.10
	Model Set 4	87.15	1.29	12.00	2.28	58.91
	Model Set 5	90.94	2.38	12.50	3.85	65.50
	Model Set 6	85.23	1.11	16.00	2.07	66.27
DTree						
	Model Set 1	97.19	3.08	6.00	4.05	52.08
	Model Set 2	97.45	4.44	6.00	4.93	52.21
	Model Set 3	92.06	2.58	16.50	4.39	59.62
	Model Set 4	76.58	1.86	39.00	3.52	61.09
	Model Set 5	80.25	1.91	35.00	3.85	60.38
	Model Set 6	76.28	1.47	36.50	2.82	56.64
Stack						
	Model Set 1	98.50	0.77	1.50	1.01	60.59
	Model Set 2	97.42	1.08	3.00	1.56	70.23
	Model Set 3	86.53	0.93	1.37	1.60	72.06
	Model Set 4	79.11	1.37	24.38	2.51	58.43
	Model Set 5	79.81	1.30	18.38	2.24	64.54
	Model Set 6	80.15	1.40	28.38	2.66	64.30
Keras						
	Model Set 1	98.94	0.00	0.00	0.00	68.82
	Model Set 2	98.03	4.10	6.00	4.75	65.23
	Model Set 3	96.54	3.16	8.00	3.93	66.40

4.5. MODEL TRANSPARENCY: VISUAL INTERPRETATION OF DECISION TREE

The best mean recall (39.00%) recorded by the DTree algorithm in this analysis was from Model Set 4), hence, we generated a representative DTree visualization (Figure 7) from this model set. The diagram is a sample visualization of the information gain analysis (with Entropy criterion) mechanism that the DTree uses in class assignment. We obtained the visualization from the 10^{th} iteration of the 10-fold StratifiedKFold cross-validation model evaluation method explored in this study. To simplify the representation and enhance understanding, we limited the tree depth to 3 (typically excluding the root node in the count), but the actual maximum depth utilized in building the DTree model in Model Set 4 is 50. In the diagram below (Figure 7), infant gender was automatically considered to be the root node because the DTree algorithm found the feature as having the highest Entropy score (1.0), hence, the rule-based random split (True/False) starts from the root node. Rules are human interpretable conditional statements used by the DTree algorithm. The internal nodes (11) are where other features and their values were tested against other partitioning rules. In total, the DTree tested 12 rules (including the root node). For the dichotomous nodes (features with binary values e.g. 0,1), the data-split rule applied by the classifier is the feature value ≤ 0.5. However, for the non-binarized categorical features (mother_blood_group for instance) with nominal values (0, 1, 2, 3 values, representing the four blood group types, ordered as {A, AB, B, O} for example), the algorithm decides and uses the best splitting point that gives the minimum error. In the same mother_blood_group example, that point was found in mother's blood group value ≤ 2.5, which corresponds to the subset {A, AB, B}. Thus, in Figure 7, the left branch corresponds to a blood type within this set, while the right branch corresponds to a blood type outside the set, and that is, blood group O. The splitting point of a feature may vary when evaluated in different subtree(s). "Samples" is the total number of

instances in each node, and the number of instances per target class ("Yes_GDHP", and "No_ GDHP") after splitting is considered by the algorithm when deciding about intermediate class assignment. A node with an even data split is regarded as an impure node (Entropy =1), hence, the near the Entropy score to 1, the more the uncertainty or impurity of the node. A node with an Entropy score nearing 0 is almost pure, and no further splitting is possible from a pure node (the purer the node, the less the information needed for its description). Therefore, the DTree algorithm continues to recursively traverse through the impure path(s) for further assessment, splitting the data at any node possible. For the ultimate or overall class assignment, the DTree classifier applies a majority rule to determine its prediction at the leaf nodes (1 leaf node is shown below, generated by the Underweight_bmi node). Majority rule involves the assignment of the most occurring class (mode) to a given instance.

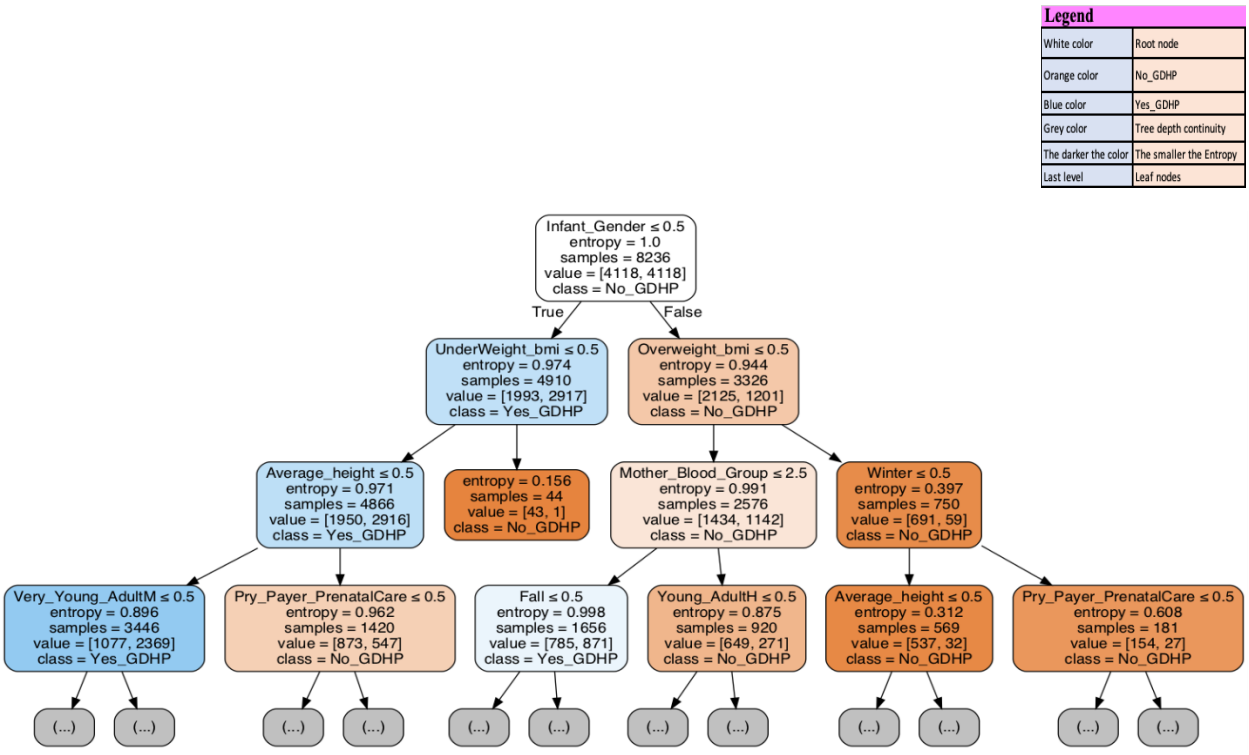


Figure 7: A sample decision tree visualization

4.6. ANSWERING RESEARCH QUESTIONS

4.6.1. RESEARCH QUESTION 1

As shown in Table 5, the best supervised machine learning tool for GDHP risk assessment was designed using the features selected by the RFECV-RF feature selection approach (Model Set 6), and not those identified by the Gen-LReg technique (Model Set 5). Therefore, there is a feature selection method capable of outperforming Gen-LReg in identifying the optimal feature subsets for modeling GDHP with the analyzed dataset, and that method is recursive feature elimination with cross validation with random forest.

4.6.2. RESEARCH QUESTION 2

Based on recall, the results (Table 5) show the optimal SML model for predicting GDHP is Support Vector Machine, thus, comparatively outperforming the Keras model on WiseSubset.

4.6.3. RESEARCH QUESTION 3

Compared to the model performance observed in Model Set 1 (Baseline); where the imbalanced WiseSubset was utilized for GDHP modeling, all other model Sets; where the SMOTE algorithm was applied, showed a demonstrable improvement (Table 5). Therefore, SMOTE was effective in resolving the class imbalance, thus, improving WiseSubset for GDHP model construction.

5. DISCUSSION OF RESULTS AND LIMITATION OF THE STUDY

5.1. DISCUSSION OF RESULTS

Considering our goal of designing a predictive SML tool (model) that optimally identifies the sub-population of pregnant women who are at risk for GDHP using the routine prenatal attributes, our model generalization and isolation decision was based on recall. The support vector machine objectively emerged as the best model to predict GDHP. The model recorded a mean recall score of 100.00%, utilizing the nine subsets of features selected by the recursive feature elimination with cross-validation with random forest as input data. This optimal model, however had a low precision score (1.06%, mean) as a tradeoff for recall, resulting in low mean F1 score (2.10%), and six negative examples were correctly labelled. Therefore, an additional screening step/further study may be needed to rule out the risk of GDHP in those falsely identified (4569) as being likely to develop the comorbidity. During the experiments, thresholding with Youden's J-Statistic was assessed, but the method did not improve the model.

Furthermore, the feature ranking results (Table 2) show six of the nine most predictive factors of GDHP identified by RFECV-RF were also among the top six selections made by at least one other feature selection approach. This lends confidence to focusing on the six risk factors namely ***healthy weight prepregnancy BMI, mother's educational status, husband's educational status, husband's occupation one year before the current pregnancy, mother's blood group, and mother's age range between 34 and 44 years.***

Healthy weight prepregnancy BMI (18.5-24.9kg/m2) ranked 5th among the 9 strongest risk factors of GDHP, and it's the 2nd of the 22 features selected by the Gen-LReg. We found the

selection and the rank of this predictor to be the most striking exposition from this study. This finding, however, agrees with prior assertions of a machine learning-based study on metabolic syndrome of the general population [19], indicating that the non-obese do sometimes develop metabolic impairments. With obesity being a traditional factor for GDM, HDP, and MetSyn, there is a tendency for its presence (or the lack thereof) to be the sole consideration for GDHP-related preventive/screening/interventional programs, thus increasing the chance of the pregnant women with a prepregnancy BMI between 18.5-24.9kg/m2 not receiving timely assessment for GDHP, a closely related disorder with metabolic syndrome. Mother's and husband's educational statuses ranked second and sixth respectively among the nine most predictive factors of GDHP. Although, the academic profile of a pregnant woman is as crucial as that of her husband in lowering or eliminating the women's risk of developing simultaneous occurrence of GDM and HDP, that of the woman has a stronger influence. Both risk factors, however, belong to the same theme (education), thus emphasizing the necessity for a joint effort of the educational and the obstetric stakeholders over this impactful obstetric morbidity. Additionally, our study makes it clear that mother's blood group (4^{th} of the 9 most predictive factors of GDHP) plays a crucial role in predicting the co-existence of GDM and HDP. This finding is consistent with literature indicating that maternal blood group is linked to HDP [48] and GDM individually [50]. Additionally, mother's age (34-44 years) ranked 8th among the 9 most influencing factors of GDHP. This finding is particularly useful for planning preventive/interventional targeted programs against the comorbidity. The 9^{th} strongest risk factor (husband's occupation one year before the current pregnancy) may be related to the family's socio-economic status (income). Lastly, the 1^{st} (inter-pregnancy interval) and the 2^{nd} (number of pregnancies plus current) of the 9

best predictors, though they were only selected by RFECV-RF, are very important obstetric factors associated with GDHP.

5.2. LIMITATION OF THE STUDY

This study has the following limitations: Both the sample size (9962) and that of its extracted sub-sample (4624) were small. The sample is also geographically limited to a hospital in WI, and the site is a member of a database in the State of WI, therefore, we do not know how well the results will generalize to other locales. Another limitation is the inability to identify the subset of features utilized by the deep learner in constructing the Keras model, but this limitation is typical of DL, hence, not a unique limitation of this research.

Lastly, based on literature, there are a few other GDM /HDP-related factors that may influence the development of GDHP, but such information was unavailable in the analyzed dataset; mother's occupation is one of such data.

6. CONCLUSION

We tested six commonly used supervised machine learning algorithms and a representative model that employs a deep learning architecture. The results showed that our study has clinical (obstetric) utility as such automated methods could be used to augment current prenatal screening to find GDHP-at-risk expectant mothers who might not otherwise draw the attention of their care provider, especially when they are 1.) primigravidas because such pregnant women would have no previous obstetric history that could prompt their care providers for MCH medical surveillance; and 2.) multigravidas with no GDHP-related past obstetric history. Therefore, the designed GDHP model, the first of its kind, would likely improve pregnancy outcomes; and ultimately be lifesaving for the gravidae and their unborn babies. The results of this research would also potentially promote an effective targeted preventive /interventional care plan; and resource distribution to the most needful (thus, saving costs) because our non-invasive ML tool additionally identified the specific and the strongest non-causal factors of GDHP.

REFERENCES

[1] L. C. Poon *et al.*, "The International Federation of Gynecology and Obstetrics (FIGO) initiative on pre-eclampsia: A pragmatic guide for first-trimester screening and prevention.," *International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics*, vol. 145 Suppl, no. Suppl 1, pp. 1–33, May 2019, doi: 10.1002/ijgo.12802.

[2] M. Sanghavi and J. D. Rutherford, "Cardiovascular physiology of pregnancy," *Circulation*, vol. 130, no. 12, pp. 1003–1008, Sep. 2014, doi: 10.1161/CIRCULATIONAHA.114.009029.

[3] A. Mito *et al.*, "Hypertensive disorders of pregnancy: A strong risk factor for subsequent hypertension 5 years after delivery," *Hypertension Research*, vol. 41, no. 2, pp. 141–146, 2018, doi: 10.1038/hr.2017.100.

[4] W. Ying, J. M. Catov, and P. Ouyang, "Hypertensive disorders of pregnancy and future maternal cardiovascular risk," *Journal of the American Heart Association*, vol. 7, no. 17, pp. 1–9, 2018, doi: 10.1161/JAHA.118.009382.

[5] J. Lu *et al.*, "Maternal Gestational Diabetes Is Associated With Offspring's Hypertension.," *American journal of hypertension*, vol. 32, no. 4, pp. 335–342, Mar. 2019, doi: 10.1093/ajh/hpz005.

[6] S. H. Riaz, M. S. Khan, A. Jawa, M. Hassan, and J. Akram, "Lack of uniformity in screening, diagnosis and management of gestational diabetes mellitus among health practitioners across major cities of Pakistan," *Pakistan Journal of Medical Sciences*, vol. 34, no. 2, pp. 300–304, 2018, doi: 10.12669/pjms.342.12213.

[7] C. W. Hockett, E. J. Bedrick, P. Zeitler, T. L. Crume, S. Daniels, and D. Dabelea, "Exposure to Diabetes in Utero Is Associated with Earlier Pubertal Timing and Faster Pubertal Growth in the Offspring: The EPOCH Study.," *The Journal of pediatrics*, vol. 206, pp. 105–112, Mar. 2019, doi: 10.1016/j.jpeds.2018.10.053.

[8] A. Rafeeinia, A. Tabandeh, S. Khajeniazi, and A. Marjani, "Metabolic Syndrome in Preeclampsia Women in Gorgan," *The Open Biochemistry Journal*, vol. 8, no. 1, pp. 94–99, 2014, doi: 10.2174/1874091x01408010094.

[9] K. Wani *et al.*, "Early-Pregnancy Metabolic Syndrome and Subsequent Incidence in Gestational Diabetes Mellitus in Arab Women," vol. 11, no. February, pp. 1–8, 2020, doi: 10.3389/fendo.2020.00098.

[10] Y. Xu, S. Shen, L. Sun, H. Yang, B. Jin, and X. Cao, "Metabolic Syndrome Risk after Gestational Diabetes: A Systematic Review and Meta-Analysis," *PLoS ONE*, vol. 9, no. 1, Jan. 2014, doi: 10.1371/JOURNAL.PONE.0087863.

[11] J. A. Grieger *et al.*, “Metabolic syndrome in pregnancy and risk for adverse pregnancy outcomes: A prospective cohort of nulliparous women,” *PLoS Medicine*, vol. 15, no. 12, pp. 1–16, 2018, doi: 10.1371/journal.pmed.1002710.

[12] C. J. Vladutiu *et al.*, “Parity and components of the metabolic syndrome among US hispanic/latina women: Results from the hispanic community health study/study of latinos,” *Circulation: Cardiovascular Quality and Outcomes*, vol. 9, no. 2_suppl_1, pp. S62–S69, 2016, doi: 10.1161/CIRCOUTCOMES.115.002464.

[13] S. A. Feresu, Y. Wang, and S. Dickinson, “Relationship between maternal obesity and prenatal, metabolic syndrome, obstetrical and perinatal complications of pregnancy in Indiana, 2008-2010,” *BMC Pregnancy and Childbirth*, vol. 15, no. 1, pp. 1–11, 2015, doi: 10.1186/s12884-015-0696-8.

[14] G. Pucci, R. Alcidi, L. Tap, F. Battista, F. Mattace-Raso, and G. Schillaci, “Sex- and gender-related prevalence, cardiovascular risk and therapeutic approach in metabolic syndrome: A review of the literature,” *Pharmacological Research*, vol. 120, pp. 34–42, 2017, doi: 10.1016/j.phrs.2017.03.008.

[15] Y. Lu *et al.*, “Metabolic mediators of the effects of body-mass index, overweight, and obesity on coronary heart disease and stroke: A pooled analysis of 97 prospective cohorts with 1·8 million participants,” *The Lancet*, vol. 383, no. 9921, pp. 970–983, 2014, doi: 10.1016/S0140-6736(13)61836-X.

[16] K. G. M. M. Alberti and P. Z. Zimmet, “Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: Diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation,” *Diabetic Medicine*, vol. 15, no. 7, pp. 539–553, 1998, doi: 10.1002/(SICI)1096-9136(199807)15:7<539::AID-DIA668>3.0.CO;2-S.

[17] K. G. M. M. Alberti, P. Zimmet, and J. Shaw, “The metabolic syndrome - A new worldwide definition,” *Lancet*, vol. 366, no. 9491, pp. 1059–1062, 2005, doi: 10.1016/S0140-6736(05)67402-8.

[18] R. J. Lipsy, “The National Cholesterol Education Program Adult Treatment Panel III guidelines.,” *Journal of managed care pharmacy : JMCP*, vol. 9, no. 1 Suppl, pp. 2–5, 2003, doi: 10.18553/jmcp.2003.9.s1.2.

[19] E. K. Choe *et al.*, “Metabolic Syndrome Prediction Using Machine Learning Models with Genetic and Clinical Information from a Nonobese Healthy Population,” *Genomics & Informatics*, vol. 16, no. 4, p. e31, 2018, doi: 10.5808/gi.2018.16.4.e31.

[20] B. Xi, D. He, Y. Hu, and D. Zhou, “Prevalence of metabolic syndrome and its influencing factors among the Chinese adults: The China Health and Nutrition Survey in 2009,” *Preventive Medicine*, vol. 57, no. 6, pp. 867–871, 2013, doi: 10.1016/j.ypmed.2013.09.023.

[21] C. M. Povel *et al.*, "Metabolic syndrome model definitions predicting type 2 diabetes and cardiovascular disease," *Diabetes Care*, vol. 36, no. 2, pp. 362–368, 2013, doi: 10.2337/dc11-2546.

[22] M. Umesawa and G. Kobashi, "Epidemiology of hypertensive disorders in pregnancy: Prevalence, risk factors, predictors and prognosis," *Hypertension Research*, vol. 40, no. 3. Nature Publishing Group, pp. 213–220, Mar. 01, 2017. doi: 10.1038/hr.2016.126.

[23] S. Saleem *et al.*, "A prospective study of maternal, fetal and neonatal deaths in low- and middle-income countries," *Bulletin of the World Health Organization*, vol. 92, no. 8, pp. 605–612, 2014, doi: 10.2471/blt.13.127464.

[24] T. Naderi, S. Foroodnia, S. Omidi, F. Samadani, and N. Nakhaee, "Incidence and Correlates of Maternal Near Miss in Southeast Iran," vol. 2015, 2015, doi: 10.1155/2015/914713.

[25] (CDC) Centers for Disease Control and Prevention Organization, "Maternal Mortality: Pregnancy Mortality Surveillance System.," 2020.

[26] L. Guariguata, U. Linnenkamp, J. Beagley, D. R. Whiting, and N. H. Cho, "Global estimates of the prevalence of hyperglycaemia in pregnancy.," *Diabetes research and clinical practice*, vol. 103, no. 2, pp. 176–185, Feb. 2014, doi: 10.1016/j.diabres.2013.11.003.

[27] A. Mito *et al.*, "Hypertensive disorders of pregnancy: a strong risk factor for subsequent hypertension 5 years after delivery.," *Hypertension research : official journal of the Japanese Society of Hypertension*, vol. 41, no. 2, pp. 141–146, Feb. 2018, doi: 10.1038/hr.2017.100.

[28] K. CK, C. S, and R. R, "Gestational diabetes and the risk of cardiovascular disease in women: a systematic review and meta-analysis," *Diabetologia*, vol. 62, no. 6, pp. 905–914, Jun. 2019, doi: 10.1007/S00125-019-4840-2.

[29] J. Porcel *et al.*, "Hypertensive disorders of pregnancy and risk of screening positive for Posttraumatic Stress Disorder: A cross-sectional study," *Pregnancy Hypertension*, vol. 3, no. 4, pp. 254–260, 2013, doi: 10.1016/j.preghy.2013.07.004.

[30] G. Martillotti *et al.*, "Preeclampsia Increased Salt Sensitivity of Ambulatory Blood Pressure in Women With a History of Severe Preeclampsia," pp. 802–808, 2013, doi: 10.1161/HYPERTENSIONAHA.113.01916.

[31] O. Odigboegwu, L. J. Pan, and P. Chatterjee, "Use of Antihypertensive Drugs During Preeclampsia," *Frontiers in Cardiovascular Medicine*, vol. 5, no. May, pp. 1–4, 2018, doi: 10.3389/fcvm.2018.00050.

[32] S. Rawal *et al.*, "Gestational diabetes mellitus and renal function: A prospective study with 9-to 16-year follow-up after pregnancy," *Diabetes Care*, vol. 41, no. 7, pp. 1378–1384, 2018, doi: 10.2337/dc17-2629.

[33] Y. Xu, S. Shen, L. Sun, H. Yang, B. Jin, and X. Cao, “Metabolic Syndrome Risk after Gestational Diabetes: A Systematic Review and Meta-Analysis,” *PLoS ONE*, vol. 9, no. 1, Jan. 2014, doi: 10.1371/JOURNAL.PONE.0087863.

[34] G. M. Maher, G. W. O’Keeffe, L. C. Kenny, P. M. Kearney, T. G. Dinan, and A. S. Khashan, “Hypertensive disorders of pregnancy and risk of neurodevelopmental disorders in the offspring: a systematic review and meta-analysis protocol.,” *BMJ open*, vol. 7, no. 10, p. e018313, Oct. 2017, doi: 10.1136/bmjopen-2017-018313.

[35] Y. Gu *et al.*, “Joint Associations of Maternal Gestational Diabetes and Hypertensive Disorders of Pregnancy With Overweight in Offspring.,” *Frontiers in endocrinology*, vol. 10, p. 645, 2019, doi: 10.3389/fendo.2019.00645.

[36] P. Damm, A. Houshmand-oeregaard, and L. Kelstrup, “Gestational diabetes mellitus and long-term consequences for mother and offspring : a view from Denmark,” *Diabetologia*, pp. 1396–1399, 2016, doi: 10.1007/s00125-016-3985-5.

[37] A. D. Martins, A. Majzoub, and A. Agawal, “Metabolic Syndrome and Male Fertility,” vol. 37, no. 2, pp. 113–127, 2019.

[38] E. Ferrazzi, T. Stampalija, L. Monasta, D. di Martino, S. Vonck, and W. Gyselaers, “Maternal hemodynamics: a method to classify hypertensive disorders of pregnancy.,” *American journal of obstetrics and gynecology*, vol. 218, no. 1, pp. 124.e1-124.e11, Jan. 2018, doi: 10.1016/j.ajog.2017.10.226.

[39] K. K, S. S, and Z. H, “Gestational diabetes mellitus and macrosomia: a literature review,” *Annals of nutrition & metabolism*, vol. 66 Suppl 2, pp. 14–20, Jun. 2015, doi: 10.1159/000371628.

[40] R. á Rogvi, J. L. Forman, P. Damm, and G. Greisen, “Women born preterm or with inappropriate weight for gestational age are at risk of subsequent gestational diabetes and pre-eclampsia,” *PLoS ONE*, vol. 7, no. 3, 2012, doi: 10.1371/journal.pone.0034001.

[41] P. Krakowiak *et al.*, “Maternal metabolic conditions and risk for autism and other neurodevelopmental disorders,” *Pediatrics*, vol. 129, no. 5, 2012, doi: 10.1542/peds.2011-2583.

[42] P. Girchenko *et al.*, “Neonatal regulatory behavior problems are predicted by maternal early pregnancy overweight and obesity: findings from the prospective PREDO Study.,” *Pediatric research*, vol. 84, no. 6, pp. 875–881, Dec. 2018, doi: 10.1038/s41390-018-0199-1.

[43] M. F. Mogos, J. L. Salemi, K. K. Spooner, B. L. McFarlin, and H. H. Salihu, “Hypertensive disorders of pregnancy and postpartum readmission in the United States: National surveillance of the revolving door,” *Journal of Hypertension*, vol. 36, no. 3, pp. 608–618, 2018, doi: 10.1097/HJH.0000000000001594.

[44] D. TM *et al.*, "The Economic Burden of Elevated Blood Glucose Levels in 2017: Diagnosed and Undiagnosed Diabetes, Gestational Diabetes Mellitus, and Prediabetes," *Diabetes care*, vol. 42, no. 9, pp. 1661–1668, Sep. 2019, doi: 10.2337/DC18-1226.

[45] S. A. Feresu, Y. Wang, and S. Dickinson, "Relationship between maternal obesity and prenatal, metabolic syndrome, obstetrical and perinatal complications of pregnancy in Indiana, 2008-2010.," *BMC pregnancy and childbirth*, vol. 15, p. 266, Oct. 2015, doi: 10.1186/s12884-015-0696-8.

[46] J. Li *et al.*, "Short Body Height and Pre-pregnancy Overweight for Increased Risk of Gestational Diabetes Mellitus: A Population-Based Cohort Study.," *Frontiers in endocrinology*, vol. 9, p. 349, 2018, doi: 10.3389/fendo.2018.00349.

[47] A. Cohen, C. F. Pieper, A. J. Brown, and L. A. Bastian, "Number of children and risk of metabolic syndrome in women.," *Journal of women's health (2002)*, vol. 15, no. 6, pp. 763–773, 2006, doi: 10.1089/jwh.2006.15.763.

[48] C. Phaloprakarn and S. Tangjitgamol, "Maternal ABO blood group and adverse pregnancy outcomes," *Journal of Perinatology*, vol. 33, no. 2, pp. 107–111, 2013, doi: 10.1038/jp.2012.73.

[49] M. Shimodaira, T. Yamasaki, and T. Nakayama, "The association of maternal ABO blood group with gestational diabetes mellitus in Japanese pregnant women," *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, vol. 10, no. 2, pp. S102–S105, 2016, doi: 10.1016/j.dsx.2016.03.003.

[50] A. Burgess, T. S. Johnson, A. Simanek, T. Bell, and S. Founds, "Maternal ABO Blood Type and Factors Associated With Preeclampsia Subtype," *Biological Research for Nursing*, vol. 21, no. 3, pp. 264–271, 2019, doi: 10.1177/1099800419833782.

[51] G. L. Booth, J. Luo, A. L. Park, D. S. Feig, R. Moineddin, and J. G. Ray, "Influence of environmental temperature on risk of gestational diabetes," *CMAJ*, vol. 189, no. 19, pp. E682–E689, May 2017, doi: 10.1503/cmaj.160839.

[52] V. Steinthorsdottir *et al.*, "Genetic predisposition to hypertension is associated with preeclampsia in European and Central Asian women," *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–14, Nov. 2020, doi: 10.1038/s41467-020-19733-6.

[53] M. Lewandowska, B. Więckowska, and S. Sajdak, "Pre-Pregnancy Obesity, Excessive Gestational Weight Gain, and the Risk of Pregnancy-Induced Hypertension and Gestational Diabetes Mellitus," *Journal of Clinical Medicine*, vol. 9, no. 6, p. 1980, Jun. 2020, doi: 10.3390/JCM9061980.

[54] (CDC) Centers for Disease Control and Prevention Organization, "About Adult BMI," 2020.

[55] A. Tandberg, K. Klungsøyr, L. B. Romundstad, and R. Skjærven, "Pre-eclampsia and assisted reproductive technologies: consequences of advanced maternal age, interbirth

intervals, new partner and smoking habits.," *BJOG : an international journal of obstetrics and gynaecology*, vol. 122, no. 7, pp. 915–922, Jun. 2015, doi: 10.1111/1471-0528.13051.

[56] C. Thomopoulos, C. Tsioufis, H. Michalopoulou, T. Makris, V. Papademetriou, and C. Stefanadis, "Assisted reproductive technology and pregnancy-related hypertensive complications: A systematic review," *Journal of Human Hypertension*, vol. 27, no. 3, pp. 148–157, 2013, doi: 10.1038/jhh.2012.13.

[57] Y. Li, X. Ren, L. He, J. Li, S. Zhang, and W. Chen, "Maternal age and the risk of gestational diabetes mellitus: A systematic review and meta-analysis of over 120 million participants," *Diabetes Research and Clinical Practice*, vol. 162, p. 108044, 2020, doi: 10.1016/j.diabres.2020.108044.

[58] A.-K. Wikström, J. Gunnarsdóttir, and S. Cnattingius, "The paternal role in pre-eclampsia and giving birth to a small for gestational age infant; a population-based cohort study.," *BMJ open*, vol. 2, no. 4, 2012, doi: 10.1136/bmjopen-2012-001178.

[59] M. Aguilar, T. Bhuket, S. Torres, B. Liu, and R. J. Wong, "Prevalence of the metabolic syndrome in the United States, 2003-2012," *Jama*, vol. 313, no. 19, pp. 1973–1974, 2015, doi: 10.1001/jama.2015.4260.

[60] L. Gregor, P. L. Remington, S. Lindberg, and D. Ehrenthal, "Prevalence of Pre-pregnancy Obesity, 2011-2014.," *WMJ : official publication of the State Medical Society of Wisconsin*, vol. 115, no. 5, pp. 228–232, Nov. 2016.

[61] L. B. Finer and M. R. Zolna, "Shifts in Intended and Unintended Pregnancies in the United States , 2001 – 2008," 2014, doi: 10.2105/AJPH.2013.301416.

[62] M. Behl *et al.*, "Evaluation of the association between maternal smoking, childhood obesity, and metabolic disorders: A national toxicology program workshop review," *Environmental Health Perspectives*, vol. 121, no. 2, pp. 170–180, 2013, doi: 10.1289/ehp.1205404.

[63] J. Wu, C. Ren, R. J. Delfino, J. Chung, M. Wilhelm, and B. Ritz, "Association between local traffic-generated air pollution and preeclampsia and preterm delivery in the South Coast Air Basin of california," *Environmental Health Perspectives*, vol. 117, no. 11, pp. 1773–1779, 2009, doi: 10.1289/ehp.0800334.

[64] M. G. Saklayen, "The Global Epidemic of the Metabolic Syndrome.," *Current hypertension reports*, vol. 20, no. 2, p. 12, Feb. 2018, doi: 10.1007/s11906-018-0812-z.

[65] B. Pratumvinit *et al.*, "Maternal vitamin d status and its related factors in pregnant women in Bangkok, Thailand," *PLoS ONE*, vol. 10, no. 7, pp. 1–14, 2015, doi: 10.1371/journal.pone.0131126.

[66] (APHA) American Public Health Association Organization, . "Call for Education and Research Into Vitamin D Deficiency/Insufficiency.," *Policy Number: 20081*, 2008.

[67] W. M. Callaghan, A. A. Creanga, and E. v. Kuklina, “Severe maternal morbidity among delivery and postpartum hospitalizations in the United States,” *Obstetrics and Gynecology*, vol. 120, no. 5, pp. 1029–1036, 2012, doi: 10.1097/AOG.0b013e31826d60c5.

[68] A. K. Amegah, M. K. Klevor, and C. L. Wagner, “Maternal Vitamin D insufficiency and risk of adverse pregnancy and birth outcomes: A systematic review and meta-analysis of longitudinal studies,” *PLoS ONE*, vol. 12, no. 3, pp. 1–22, 2017, doi: 10.1371/journal.pone.0173605.

[69] R. Oza-Frank, I. Chertok, and A. Bartley, “Differences in breast-feeding initiation and continuation by maternal diabetes status.,” *Public health nutrition*, vol. 18, no. 4, pp. 727–735, Mar. 2015, doi: 10.1017/S1368980014000792.

[70] E. E. Petersen, N. L. Davis, D. Goodman, S. Cox, C. Syverson, and K. Seed, “Racial / Ethnic Disparities in Pregnancy-Related Deaths — United States , 2007 – 2016,” 2016.

[71] R. Retnakaran *et al.*, “Fetal Sex and Maternal Risk of Gestational Diabetes Mellitus : The Impact of Having a Boy,” vol. 38, no. May, pp. 844–851, 2015, doi: 10.2337/dc14-2551.

[72] J. H. Jhee *et al.*, “Prediction model development of late-onset preeclampsia using machine learning-based methods,” *PLoS ONE*, vol. 14, no. 8, pp. 1–12, 2019, doi: 10.1371/journal.pone.0221202.

[73] T. R. Gaillard and T. R. Gaillard, “The Metabolic Syndrome and its Components in African-American women : emerging Trends and implications,” no. January, 2018, doi: 10.3389/fendo.2017.00383.

[74] W. Kim, S. K. Park, and Y. L. Kim, “Gestational diabetes mellitus diagnosed at 24 to 28 weeks of gestation in older and obese women: Is it too late?,” *PLoS ONE*, vol. 14, no. 12, pp. 1–16, 2019, doi: 10.1371/journal.pone.0225955.

[75] V. Bolón-Canedo and A. Alonso-Betanzos, “Ensembles for feature selection: A review and future trends,” *Information Fusion*, vol. 52, pp. 1–12, Dec. 2019, doi: 10.1016/J.INFFUS.2018.11.008.

[76] R. Blagus and L. Lusa, “SMOTE for high-dimensional class-imbalanced data,” *BMC Bioinformatics 2013 14:1*, vol. 14, no. 1, pp. 1–16, Mar. 2013, doi: 10.1186/1471-2105-14-106.

[77] X. Yang, “Identification of risk genes associated with myocardial infarction based on the recursive feature elimination algorithm and support vector machine classifier,” *Molecular Medicine Reports*, vol. 17, no. 1, pp. 1555–1560, 2018, doi: 10.3892/mmr.2017.8044.

[78] V. R. Elgin Christo, H. Khanna Nehemiah, B. Minu, and A. Kannan, “Correlation-based ensemble feature selection using bioinspired algorithms and classification using backpropagation neural network,” *Computational and Mathematical Methods in Medicine*, vol. 2019, 2019, doi: 10.1155/2019/7398307.

[79] T. Santhanam and M. S. Padmavathi, “Application of K-Means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis,” *Procedia Computer Science*, vol. 47, no. C, pp. 76–83, 2015, doi: 10.1016/j.procs.2015.03.185.

[80] B. Farran, A. M. Channanath, K. Behbehani, and T. A. Thanaraj, “Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: Machine-learning algorithms and validation using national health data from Kuwait-a cohort study,” *BMJ Open*, vol. 3, no. 5, pp. 1–10, 2013, doi: 10.1136/bmjopen-2012-002457.

[81] Y. Ye, Y. Xiong, Q. Zhou, J. Wu, X. Li, and X. Xiao, “Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study.,” *Journal of diabetes research*, vol. 2020, p. 4168340, 2020, doi: 10.1155/2020/4168340.

[82] H. Sufriyana, Y. W. Wu, and E. C. Y. Su, “Artificial intelligence-assisted prediction of preeclampsia: Development and external validation of a nationwide health insurance dataset of the BPJS Kesehatan in Indonesia,” *EBioMedicine*, vol. 54, 2020, doi: 10.1016/j.ebiom.2020.102710.

[83] R. J. Kate, R. M. Perez, D. Mazumdar, K. S. Pasupathy, and V. Nilakantan, “Prediction and detection models for acute kidney injury in hospitalized older adults,” *BMC Medical Informatics and Decision Making*, vol. 16, no. 1, pp. 1–11, 2016, doi: 10.1186/s12911-016-0277-4.

[84] B. Zhang, L. Lu, and J. Hou, “A comparison of logistic regression, random forest models in predicting the risk of diabetes,” *ACM International Conference Proceeding Series*, pp. 231–234, 2019, doi: 10.1145/3364836.3364882.

[85] A. Vellido, J. D. Martín-Guerrero, and P. J. G. Lisboa, “Making machine learning models interpretable,” *ESANN 2012 proceedings, 20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, no. April, pp. 163–172, 2012.

[86] K. Doubleday, H. Zhou, H. Fu, and J. Zhou, “An Algorithm for Generating Individualized Treatment Decision Trees and Random Forests.,” *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, vol. 27, no. 4, pp. 849–860, 2018, doi: 10.1080/10618600.2018.1451337.

[87] A. Ahmad, A. Mustapha, E. D. Zahadi, N. Masah, and N. Y. Yahaya, “Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus,” 2011.

[88] C. Galaviz-Hernandez, M. Sosa-Macias, E. Teran, J. E. Garcia-Ortiz, and B. P. Lazalde-Ramos, “Paternal Determinants in Preeclampsia.,” *Frontiers in physiology*, vol. 9, p. 1870, 2018, doi: 10.3389/fphys.2018.01870.

[89] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Medicine*, vol. 17, no. 1. BioMed Central Ltd., Oct. 29, 2019. doi: 10.1186/s12916-019-1426-2.

[90] Y. Du *et al.*, "Application of ultrasound-based radiomics technology in fetal lung texture analysis in pregnancies complicated by gestational diabetes or pre-eclampsia.," *Ultrasound in obstetrics & gynecology : the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*, Apr. 2020, doi: 10.1002/uog.22037.

[91] S. Kong and Y. S. Cho, "Identification of female-specific genetic variants for metabolic syndrome and its component traits to improve the prediction of metabolic syndrome in females," *BMC Medical Genetics*, vol. 20, no. 1, pp. 1–13, 2019, doi: 10.1186/s12881-019-0830-y.

[92] K. K. Ryckman, K. S. Borowski, N. I. Parikh, and A. F. Saftlas, "Pregnancy Complications and the Risk of Metabolic Syndrome for the Offspring," *Current Cardiovascular Risk Reports*, vol. 7, no. 3, pp. 217–223, 2013, doi: 10.1007/s12170-013-0308-y.

[93] L. J. Li *et al.*, "Effect of gestational diabetes and hypertensive disorders of pregnancy on postpartum cardiometabolic risk," *Endocrine Connections*, vol. 7, no. 3, pp. 433–442, 2018, doi: 10.1530/EC-17-0359.

[94] W. Cao *et al.*, "Maternal lipids, BMI and IL-17/IL-35 imbalance in concurrent gestational diabetes mellitus and preeclampsia," *Experimental and Therapeutic Medicine*, vol. 16, no. 1, pp. 427–435, 2018, doi: 10.3892/etm.2018.6144.

[95] "Centers for Disease Control and Prevention, About Adult BMI," *Division of Nutrition, Physical Activity, and Obesity, National Center for Chronic Disease Prevention and Health Promotion*, p. 2020, 2017, [Online]. Available: https://www.cdc.gov/healthyweight/assessing/bmi/adult_bmi/index.html#

APPENDICES

APPENDIX A: THE STARTING FIELD LIST IN WISESAMPLE

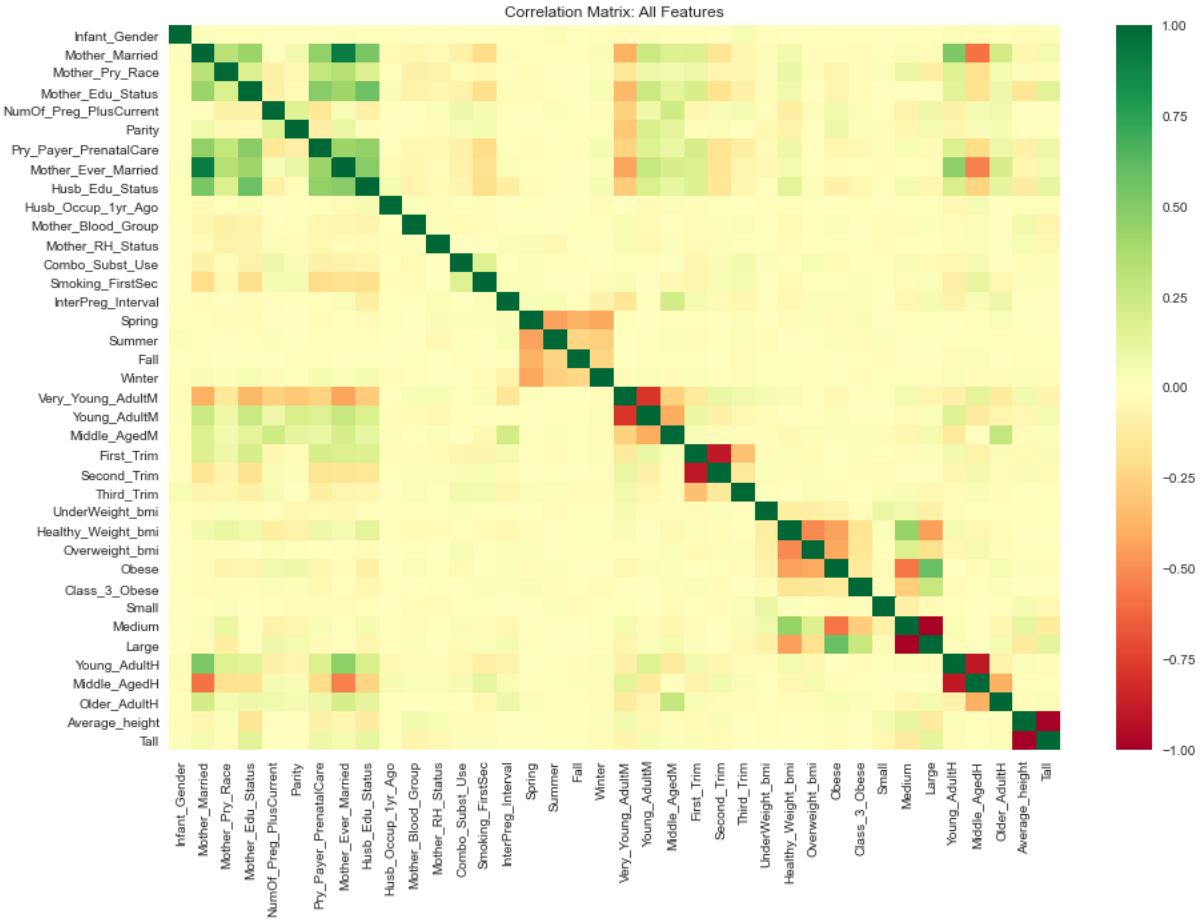
The Starting 79 fields in WiseSample					
Serial Number	**Feature Name**	**Serial Number**	**Feature Name**	**Serial Number**	**Feature Name**
1	Infant_Gender	31	GDM_this_Pregnancy	55	Anemia
2	Mother_Married	32	GDM_this_Preg_Insulin_Used	56	Mother_Calc_Age
2	Pat_Signature	33	VRecord_HTN_Prepreg	57	Mother_Reported_Age
4	Mother_Pry_Race	34	VRecord_GHTN	58	Mother_Ever_Married
5	Mother_Edu_Status	35	Eclampsia	59	Bio_Father_IsLegal_Husb
5	Previous_Pregnancies	36	Multiple_Gest	60	Husb_Calc_Age
7	NumOf_Preg_PlusCurrent	37	Cocaine_Use	61	Husb_Reported_Age
8	Parity	38	Methamphetamine_Use	62	Husb_Edu_Status
9	Year_Last_Birth	39	Heroin_Use	63	Husb_Occup_1yr_Ago
10	Pry_Payer_PrenatalCare	40	Hallucinogens_Use	64	Father_Calc_Age
11	Prepreg_Weight_lbs	41	Marijuana_Use	65	Father_Reported_Age
12	Prepreg_BMI	42	Other_Drugs_Used	66	Father_Pry_Race
13	Mother_Weight_At_Delivery_lbs	43	Antihypertensivesants	67	Father_Edu_Status
14	Mother_Weight_Change	44	HELLP_Syndrome	68	Father_Occup_1yr_Ago
15	Gest_Week_PNCare_Began	45	Antihypertensives	69	Mother_Height_Feet
16	Smoking	46	Hysterotomy_Hysterectomy	70	Mother_Height_Inches
17	Mother_Lives_with_Smoker	47	Number_Of_Fetuses_Delivered	71	Date_First_PNCVisit
18	Alcohol_Use_During_Preg	48	Live_Births	72	Mother_Blood_Group
19	CS_PreviousPreg	49	Amphetamines	73	Mother_RH_Status
20	Preterm_Birth_PreviousPreg	50	Barbiturates	74	PE_without_Severe_Feats
21	LBW_PreviousPreg	51	Cocaine	75	PE_with_Severe_Feats
22	VLBW_PreviousPreg	52	Marijuana	76	Substance_Abuse_Rohypnol
23	ELBW_PreviousPreg	53	Opiates	77	HTN_in_Preg_Chronic
24	SGA_PreviousPreg	54	Other_Substances	78	GHTN_Mat_Record
25	IUGR_PreviousPreg			79	Previous_Preeclampsia
26	Macrosomia_PreviousPreg				
27	Infertility_Treatment				
28	Assisted_Reproduction_Tech				
29	Fertility_Enhancing_Drugs				
30	Diab_Prepreg				

APPENDIX B: FINAL FIELD LIST WITH THEIR VALUES, AND THE 9 OPTIMAL FEATURES

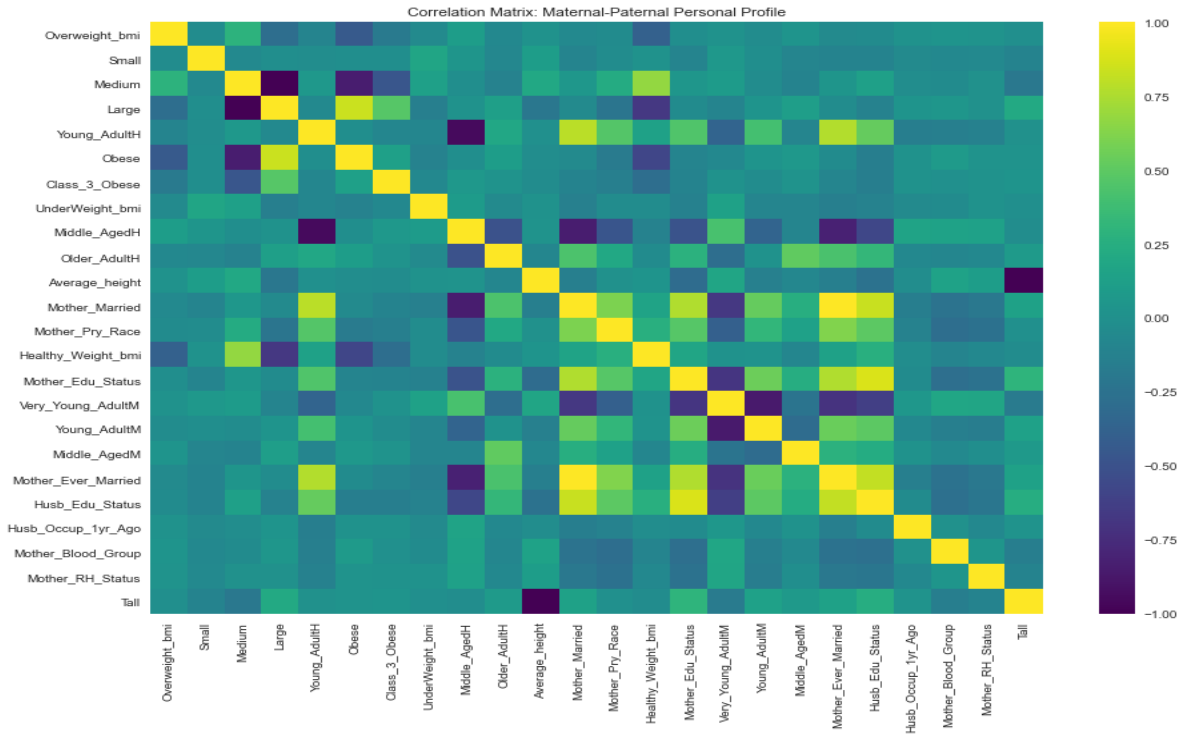
Key:
Rank: Non-normalized ranking (in descending order) of the 9 strongest predictors of GDHP.
Feature names written in red ink are the 24 new features (with their corresponding values).
Working definitions were utilized for the 24 new features where applicable.

Serial Number	Feature Names (with full meaning)	Value/Range of Values	Rank
1	Infant_Gender	Male/Female	
2	Mother_Married	Yes/No	
3	Mother_Pry_Race *(Pry= Primary)*	Black, White, Others	
4	Mother_Edu_Status *("Edu"= Education)*	(7 unique values): 8Th_Grd_Or_Less, 9Th-12Th Grd_Nodiploma, High School Degree Or Ged, Associate, Bachelor, Master, Doctorate_Or Prof_Deg	2
5	NumOf_Preg_PlusCurrent *("Preg"= Pregnancy)*	10 unique values: 2-11	3
6	Parity	Nulliparous/multiparous	
7	Pry_Payer_PrenatalCare *("Pry"= Primary)*	Medicaid, Private, Others	
8	Mother_Ever_Married	Yes/No	
9	Husb_Edu_Status *("Husb"= Husband)*	(7 unique values): 8Th_Grd_Or_Less, 9Th-12Th Grd_Nodiploma, High School Degree Or Ged, Associate, Bachelor, Master, Doctorate_Or Prof_Deg	6
10	Husb_Occup_1yr_Ago *("Occup"= Occupation)*	1190 unique values	9
11	Mother_Blood_Group	A, AB, O, and B	4
12	Mother_RH_Status *("RH"= Rhesus)*	Positive, Negative	
13	Combo_Subst_Use *(Substance Use)*	Yes/No	
14	Smoking_FirstSec *("FirstSec" = First/Secondhand)*	Yes/No	
15	InterPreg_Interval	20 unique values: 0-19 years	1
16	Spring *(Season of the first prenatal care visit)*	Jan-April	
17	Summer *(Season of the first prenatal care visit)*	April-July	
18	Fall *(Season of the first prenatal care visit)*	July-October	
19	Winter *(Season of the first prenatal care visit)*	October-December	
20	Very_Young_AdultM *("M"= Mother)*	14-24 years	
21	Young_AdultM *("M"= Mother)*	24-34 years	
22	Middle_AgedM *("M"= Mother)*	34-44 years	8
23	First_Trim *("Trim"= Trimester)*	0-12 gestational weeks	
24	Second_Trim *("Trim"= Trimester)*	12-25 gestational weeks	
25	Third_Trim *("Trim"= Trimester)*	25-41 gestational weeks	
28	Small *(Mother's prepregnancy weight)*	0-96 lbs	
29	Medium *(Mother's prepregnancy weight)*	96-180 lbs	7
30	Large *((Mother's prepregnancy weight)*	180-301 lbs	
31	Young_AdultH ("H"= Husband)	Approx: 20-30 years	
32	Middle_AgedH *("H"= Husband)*	30-40 years	
33	Older_AdultH *("H"= Husband)*	40-50 years	
34	Average_height	4.10-5.40 ft.	
35	Tall	5.40-6.70 ft	
26	Healthy_Weight_bmi *(Mother's prepregnancy BMI)*	18.5-24.9 kg/m2	5
27	Overweight_bmi *(Mother's prepregnancy BMI)*	24.9-29.9 kg/m2	
36	Obese *(Mother's prepregnancy BMI)*	29.9-39.9 kg/m2	
37	Class_3_Obese *(Mother's prepregnancy BMI)*	39.9-100 kg/m2	
38	UnderWeight_bmi	0-18.5 kg/m2	

APPENDIX C: CORRELATION MATRIX (THE ENTIRE 38 FEATURES)

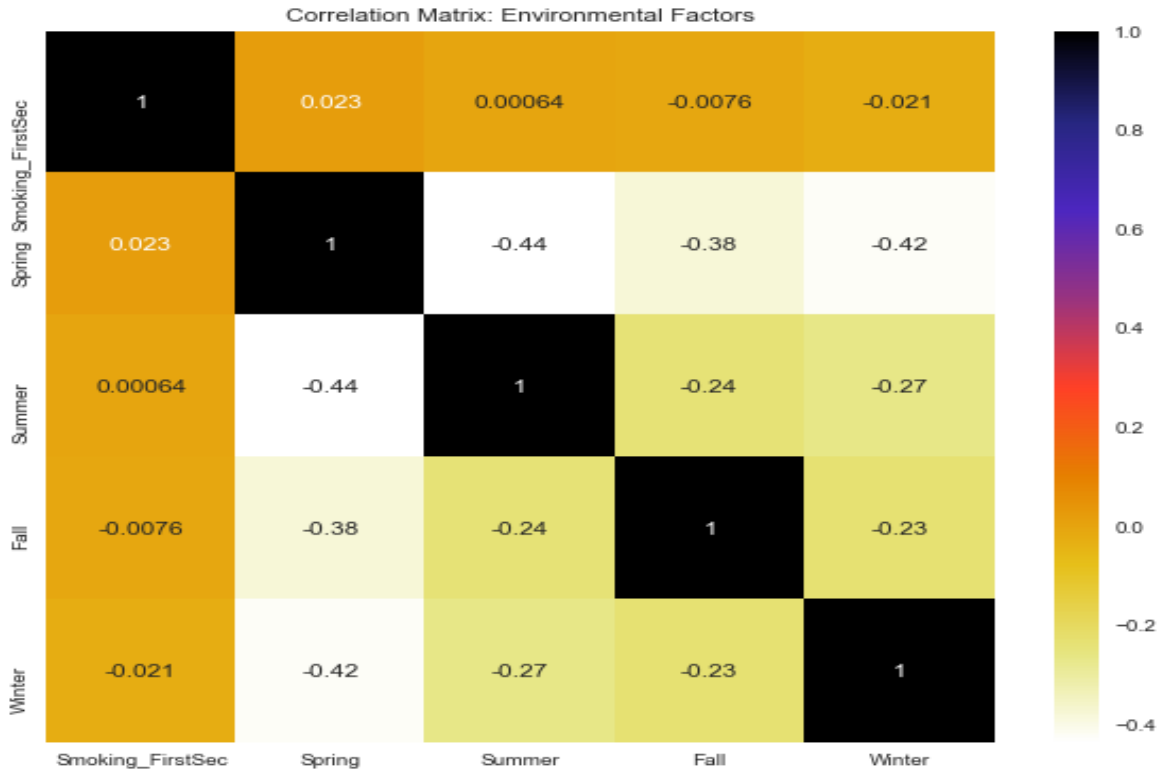


APPENDIX D: CORRELATION MATRIX (MATERNAL-PATERNAL PERSONAL PROFILE THEME)

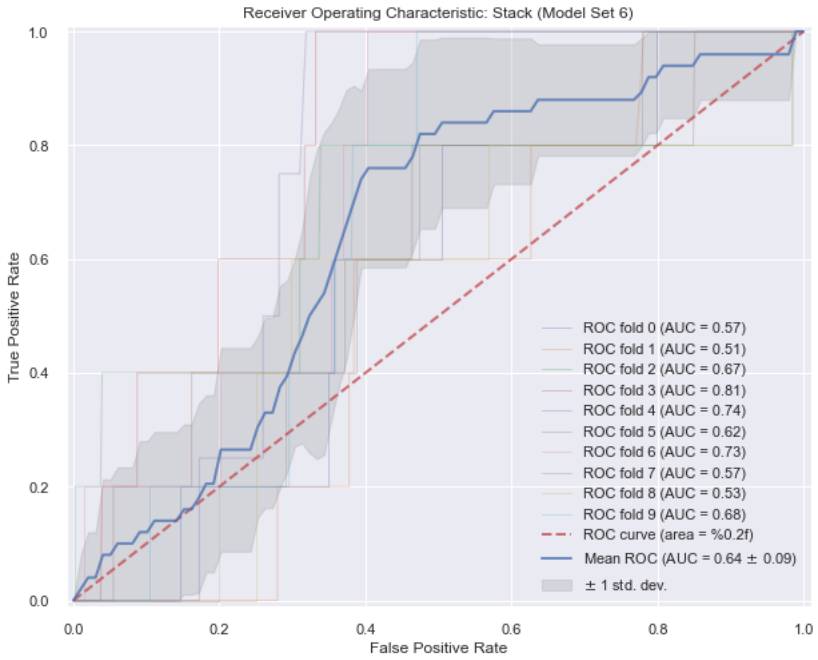


APPENDIX E: CORRELATION MATRIX (OBSTETRIC THEME)

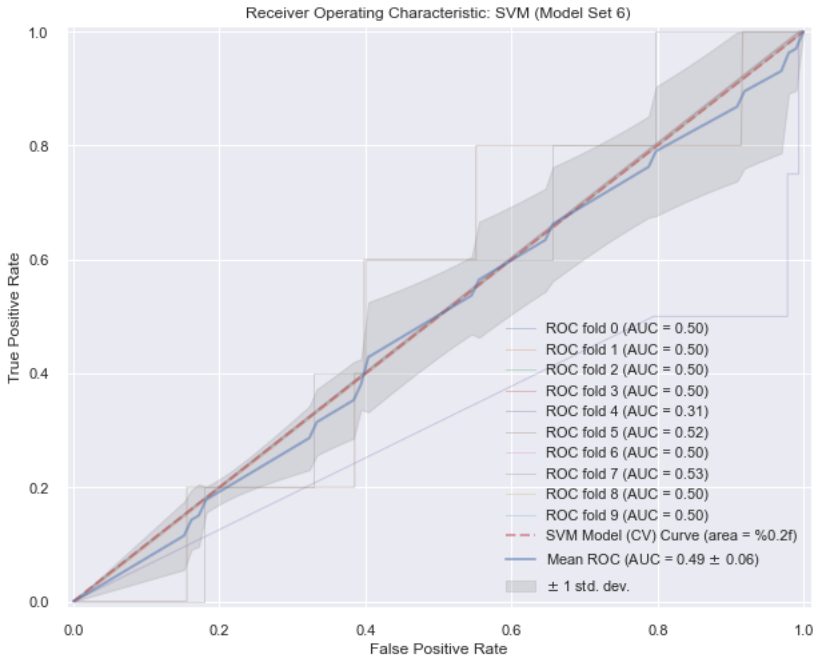
APPENDIX F: CORRELATION MATRIX (ENVIRONMENTAL THEME)



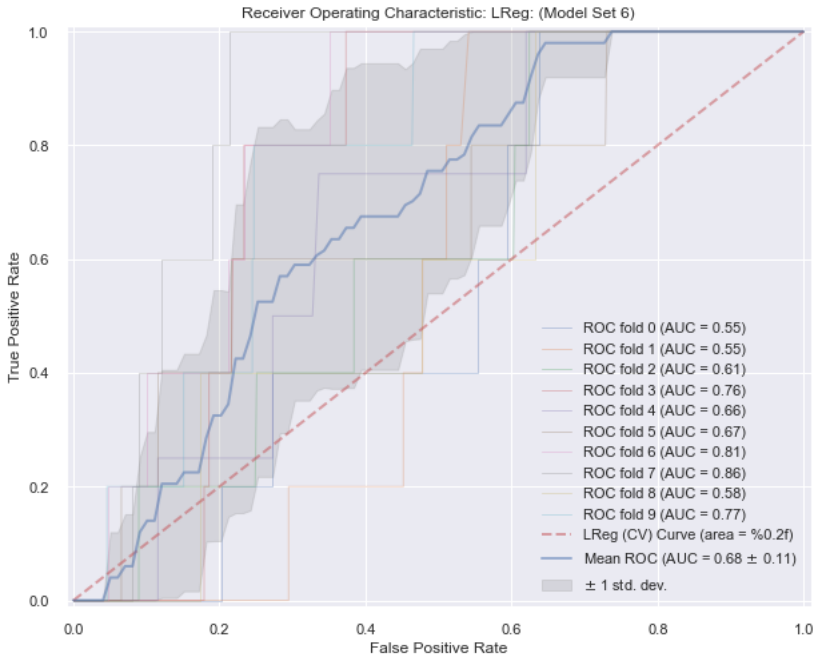
APPENDIX G: ROC OF THE STACKINGCLASSIFIER (MODEL SET 6)



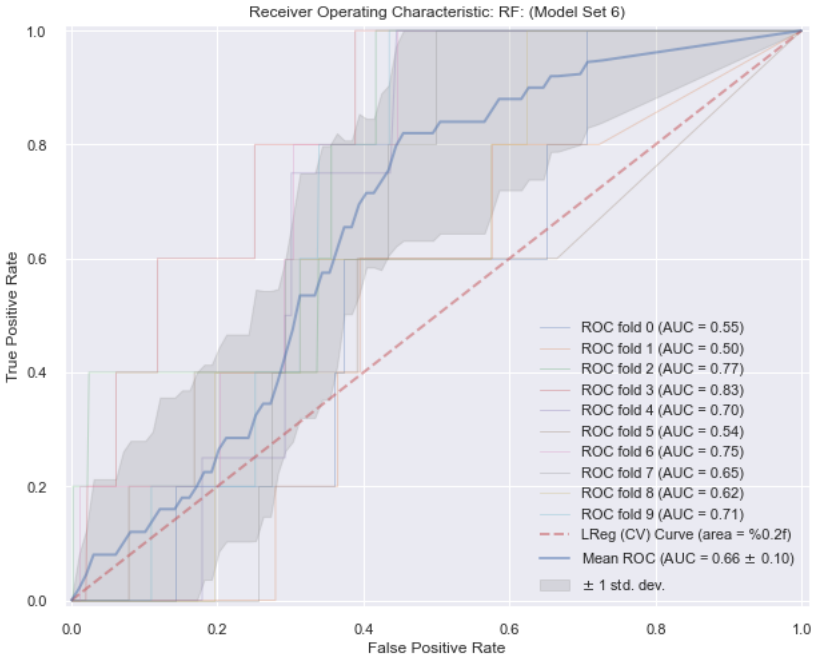
APPENDIX H: ROC OF THE SUPPORT VECTOR MACHINE (MODEL SET 6)



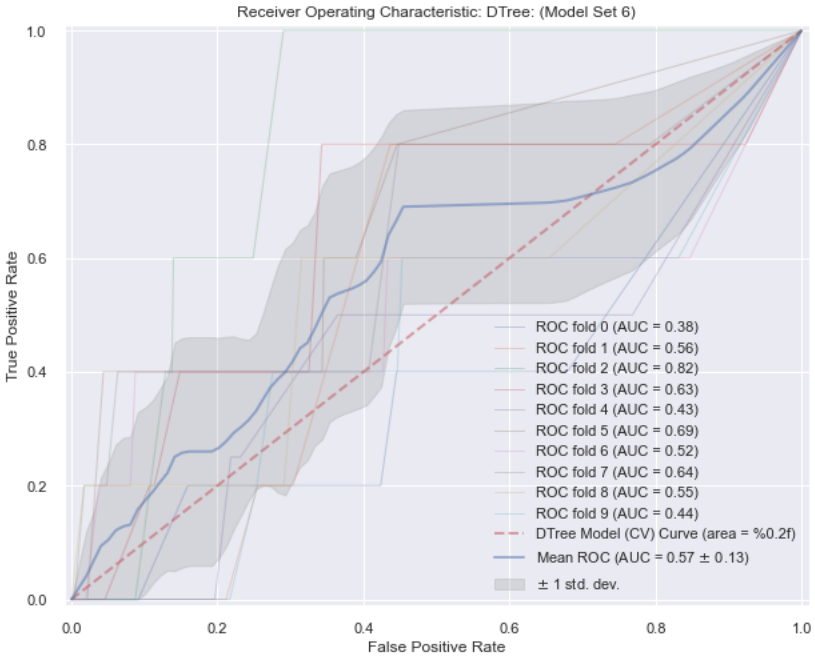
APPENDIX I: ROC OF THE LOGISTIC REGRESSION (MODEL SET 6)



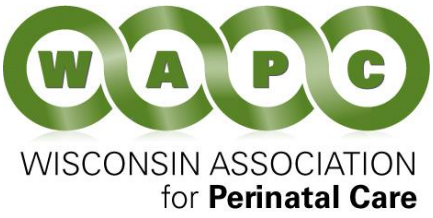
APPENDIX J: ROC OF THE RANDOM FOREST (MODEL SET 6)



APPENDIX K: ROC OF THE DECISION TREE (MODEL SET 6)



Acknowledgement: We would like to acknowledge the following organizations for the use of PeriData.Net®, Wisconsin Association for Perinatal Care, Ancilla Partners, Inc., and the hospitals that contributed their data for this project.



CURRICULUM VITAE

MARY EJIWALE

Education

PhD: Biomedical & Health Informatics, University of Wisconsin, Milwaukee (UWM). Date: *2016 -2021*

Related Core Coursework Completed:

Predictive Analytics in Healthcare, Artificial Intelligence, Algorithm Design and Analysis, Introduction to Databases, Human Pathophysiology, Introduction to Health Care Informatics, Multivariate Statistics for Healthcare, Text Retrieval and its Application in Biomedicine, Outcomes and Quality Management**,** Engineering Statistical Analysis, Intelligent User Interfaces, Biostatistics for Nursing Practice, Essentials Programming in Health Informatics, Natural Language Processing, Biomedical and Healthcare Terminology and Ontology, Maternal and Child Health Foundations, Policy and Practice; Big Data Analytics in Healthcare.
Dissertation Topic: **Prediction of Concurrent Hypertensive Disorder in Pregnancy and Gestational Diabetes Using Machine Learning Techniques.**

Master's: Health Informatics and Information Management, School of Health Related Professions, University of Mississippi Medical Center (UMMC), Jackson, MS. Date: *2012-2014*

Master's Practicum Topic**: The Influence of Computerized Physician Order Entry (CPOE) on Medication Error in Ambulatory Settings**

Non-degree (Computer Science: Undergraduate core courses only): Jackson State University, Jackson, MS. Date: *2015–2016*

Related Computer Science Core Coursework Taken:

Programming Fundamentals- Java (Lecture and Lab), Data Structures and Algorithm (Lecture and Lab), Object-Oriented Programming (Lecture and Lab), Discrete Structures for Computer Science, Geographic Information Systems.

BSc/Ed: Health Education, Lagos State University, Lagos, Nigeria. Date: *1999-2002*

Transfer Science Courses: Hinds Community College, Raymond, MS. Date: *2009-2012*

Registered Midwife (RM): Midwifery School, Baptist Medical Center, Saki, Nigeria. Date: *1994-1995*

Registered Nurse (RN): Oyo State School of Nursing (now, Oyo State College of Nursing and Midwifery), Ibadan, Nigeria. Date: *1990-1993*

Peer Reviewed Papers

A.A. Onoja, **M.O. Ejiwale**, Z. Yu, Amal Mitra, A. Rewane, B. Amao, G. Bertarelli, C. Giusti, M. Nanni. Transactions on Networks and Communications, 9(2), 2021. DOI: 10.14738/tnc.92.9760.

***Three papers are on the way*

Other Research/ Presentation

- Doctoral Seminar: Opioid Use in Pregnancy at the Institute for Health and Equity, Medical College of Wisconsin. Date: *2019*
- Poster: Collaborative Research Engine; Building a Graph Database to Promote Team Science among the Medical College of Wisconsin Faculty: Medical College of Wisconsin. Date: *2017.*
- Poster: m2Health Application, A Consumer Health Informatics Tool for Evidence-Based Self-Management Care Against Postpartum Infection: Annual Student Research Poster Competition, College of Engineering, UWM. Date: *2017*
- Poster: m2Health Application, A Consumer Health Informatics Tool for Evidence-Based Self-Management Care Against Postpartum Infection: College of Health Sciences, University of Wisconsin, Milwaukee, Date: *2017*
- Poster: Geo-coding Obesity; Food Outlets Location as a Factor for Human Size. JSU, Jackson, MS. Date: *2016*

Short Course

Role Delineation Course: Mississippi Board of Nursing, Flowood, MS. Date: *2011*

Community Health Advocate at the UMMC, Jackson, MS. Date: 2*012*

Certification/ Licensure

Registered Health Information Administrator (RHIA). Date: *2014*

Certified Associate of Health Information Management Systems (CAHIMS)- Healthcare and Information Management Systems Society (HIMSS). Date: *2013*

RN: Nursing and Midwifery Council of Nigeria. Date: *1993.*

RM: Nursing and Midwifery Council of Nigeria. Date: *1995*

Award

Graduate Area of National Need (GAANN) Fellowship. Date: *2016-2020*.

Chancellor's Award- UWM. Date: *2017/2018* Excellence in Coding (ICD 9/10 & CPT), UMMC, Jackson, MS. Date: *2012*

Dean's Award- Hinds Community College, Raymond, MS. Date: *2011*

Tutoring

ICD 9, ICD 10 and CPT Medical Coding, School of Health Related Professions, UMMC, Jackson, MS. Date: *2013-2014*

Teaching Assistant	Discrete Information Structures, Computer Science Department, UWM *2017/18/19* Introduction to Database Systems, Computer Science Department, UWM *2018-2021*
Research Assistant	Massie Chair Cybersecurity, JSU, Jackson, MS. Date: *2015- 2016*
Work Experience	Nursing Assistant: UMMC, Jackson, MS. Date: *2010- 2012* Registered Nurse, Massey Children Hospital Lagos, Nigeria, Date: *1998- 2004* Registered Nurse/Midwife, Ajayi Med. Center Lagos, Nigeria. Date: *1996-1998* Registered Nurse/ Midwife, Yombo Hospital, Lagos, Nigeria, Date: *1995-1996*
Mentoring	Social Entrepreneur, Justice & Equity Compact (SEJEC), UWM. Date: *2018* Summer Engineering Experience for Kids (SEEK). Cite: Jackson, MS. Organization: NSBE. Date: *2015/2016* Coding Hackathon for Children, Jackson, MS. Date: *2015* YesWeCode Hackathon *(*TECH JXN). Date: *2015*
Internship	Data Science: 500 STARS, Clinical and Translational Science Institute, Medical College of Wisconsin, Wauwatosa, WI. Date: *2017/2019* ICD 10 Medical Specialty Training, UMMC, Jackson, MS. Date: *2014* Data Analysis (HCAHPS data, National Research Corporation): Quality Improvement Department, UMMC, Jackson, MS. Date: *2014*
Professional Organization	American Health Information Management Association (AHIMA) Healthcare information & Management Systems Society (HIMSS) American Medical Informatics Association (AMIA)
Community Service	UWM College of Engineering Booth, Maker Fair Milwaukee. Date: 2*019* Motivating the Elementary School Students for Computing, UWM. Date: *2018* Career Day: Parkview Elementary School, Milwaukee. Date: 20*19*Medical Information Release Officer (Volunteer) Jackson Free Clinic, MS. Date: *2013-2016* Culture Day, Boys Scout of America, Clinton, MS. Date: *2015* Poverty Simulation Event, UMMC, Jackson, MS. Date: *2013* Jackson Free Clinic Fundraising Event, UMMC, Jackson, MS. Date: *2013*
International Voluntary Services	Donation of two oxygen concentrators by GLoSCA Inc. to the Baptist Medical Center, Saki, Nigeria. Grant Written by Mary Ejiwale Awarding Organization: The Build-A-Bear Foundation, USA. Date: *2010* Breast Milk Donation- International Breast Milk Project Bank, USA. Date: *2007* Fundraising- Malawi Project, Connecting the Dots, Jackson, MS. Date: *2013*
Capabilities & Skills	Programming Language- Python and Java. Statistical software- SPSS. Data Science: Machine Learning (including Deep Learning), Natural Language Processing, Big

Data analysis with Apache Spark; Search Platforms- Apache Solr and Apache Lucene. Relational Database- Structural Query Language (SQL)- Oracle and MySQL. Graph Database- Neo4j, Application Development (Android Studio). Chatbots development with IBM Watson (AI). Others are: Grant writing, Policy Brief, Microsoft Office Suite, and ArcGIS.

References

Susan McRoy

Computer Science Department

College of Engineering and Applied Sciences

University of Wisconsin- Milwaukee

Email: mcroy@uwm.edu

Phone: 414-229-6695

Folajimi Yetunde

Computer Science Department

Wentworth Institute of Technology

550 Huntington Ave, Boston, MA 02115

Email: folajimy@wit.edu

Phone: 617-989-4249

Lisa Morton

Department of Health Informatics and Information Management

School of Health-Related Professions (SHRP)

University of Mississippi Medical Center

2500, North State St, Jackson, MS 39216

Phone: 601- 984-6305

E-mail: memorton@umc.edu