A HYBRID APPROACH OF FEATURE SELECTION AND K-NEAREST NEIGHBOR FOR HANDLING HEALTHCARE PROSPECTIVE MISSING DATA

BY

RAGHAD ALKHAWALDEH

B.S., Mechanical Engineering, Jordan University of Science and Technology, 2014
M.S., Industrial and Systems Engineering, Binghamton University, 2016

DISSERTATION

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Industrial and Systems Engineering
in the Graduate School of
Binghamton University
State University of New York
2023

Accepted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Industrial and Systems Engineering
in the Graduate School of
Binghamton University
State University of New York
2023

Oct 2, 2023

Dr. Sarah S. Lam, Chair and Faculty Advisor
Department of Systems Science and Industrial Engineering, Binghamton University

Dr. Mohammad Khasawneh, Committee Member
Department of Systems Science and Industrial Engineering, Binghamton University

Dr. Bing Si, Committee Member
Department of Systems Science and Industrial Engineering, Binghamton University

Dr. Daehan Won, Committee Member
Department of Systems Science and Industrial Engineering, Binghamton University

Dr. Kyoung-Don Kang Outside Examiner
Department of Computer Science, Binghamton University

Abstract

The electronic health record system generates an enormous amount of data that can be utilized in healthcare predictive analytics. Although there has been a steady increase in the number of publications that focused on healthcare predictive analytics, publications that focused on prospective predictive models remain to be limited. Most of the proposed models in the literature are retrospective models that delivered insights about patient populations. There is a great value in applying predictive analytics prospectively in the healthcare space, such as preventing patient harm, improving care outcomes, and reducing costs and wastes. However, the deployment of predictive models prospectively continues to face barriers and some limitations. Missing data prospectively is one of these barriers that is addressed in this research. When models, trained and validated on retrospective datasets, are applied on prospective datasets such as active patient records, they can fail or produce different results due to missing or delayed model inputs. Different data imputation methods can be applied on prospective datasets when predictive models are deployed. However, these methods impact model prediction accuracy and can cause bias in model outcomes.

This dissertation proposes two hybrid approaches that combine wrapper and filter feature selection methods with K-Nearest Neighbor (KNN) imputation to impute prospectively missing variables. The first approach is called Class Driven Feature Selection KNN (CDFS-KNN). The second approach is called Missing Variable Driven Feature Selection KNN (MVDFS-KNN). CDFS-KNN and MVDFS-KNN utilize KNN imputation with a modified search space that is defined by wrapper feature selection and filter feature selection methods, respectively. The performance of the proposed methods was compared with two common imputation methods, which include mean imputation and traditional KNN imputation. Different levels of missingness (from one to five missing variables) were simulated and evaluated to measure the robustness of the two proposed approaches. Multiple benchmark datasets with different number of instances, dimensionality, and class imbalance were utilized to demonstrate the generalizability of the two proposed approaches.

The results of this study show that as the number of missing variables increased, both CDFS-KNN and MVDFS-KNN imputation resulted in the smallest degradation in models' performance when compared to KNN imputation and mean imputation. The average drop in F1 scores for CDFS-KNN and MVDFS-KNN were 0.11-4.30% and 0.17-4.44%, respectively, whereas the average drop in F1 scores for KNN imputation and mean imputation were 0.19%-5.00% and 0.32%-27.59%, respectively. It was observed that the performance of predictive models on prospective data deteriorates as the number of imputed missing increases from one to five, regardless of the data imputation method used.

The two proposed approaches will enable the use of any variable available retrospectively in constructing predictive models, without limitations to only the variables available prospectively. It will also allow for deploying models prospectively that can achieve similar performance to models that are based on retrospective testing and validation. This would potentially support the clinical decision-making process, applying early interventions, and achieving measurable improvements in healthcare outcomes.

To my parents, my husband, and my daughter Tulipa

Thank you for believing in me and providing me with your unwavering support and presence throughout the course of my doctoral pursuit.

ACKNOWLEDGEMENTS

I would like to express my appreciation to my advisor Dr. Sarah Lam for her continuous support during my doctoral journey. She is an exceptional advisor and mentor; She has consistently offered me innovative insights and motivated me to engage in thorough analysis and extensive research. I am always grateful for her guidance, motivation, and encouragement. I am profoundly appreciative for the opportunity to collaborate and acquire knowledge under the tutelage of such an exceptionally accomplished professor.

I owe a deep sense of gratitude to Dr. Mohammad T. Khasawneh for his continuous support and encouragement during my Ph.D. journey. He has perpetually instilled in me a strong work ethic and served as a motivating force to propel me closer to the finish line.

I wish to express my heartfelt gratitude to my loving husband Ashraf, whose unwavering love and steadfast support guided me through the entirety of my Ph.D. journey. He remained a constant source of strength during both the peaks and the valleys of this pursuit, consistently motivating me to strive for excellence.

I wish to extend my sincere gratitude to my mother, who has consistently stood by my side, providing support and encouragement to help me achieve my best. Words cannot adequately convey the depth of my appreciation, as I genuinely believe that I would not have accomplished this without her constant presence and guidance.

I would like to express my heartfelt gratitude to my sister, Diala, who have consistently provided her encouragement and support, despite the physical distances that separated us.

Finally, I would like to extend my gratitude to my friend Rawan Najar. She has been a constant source of encouragement, offering invaluable advice and a compassionate ear to listen to my thoughts. I genuinely appreciate the guidance and care she has generously provided me with.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ANOVA	Analysis of Variance
AUC	Area Under Curve
BCW	Breast Cancer Wisconsin
CC	Cervical Cancer
CDFS	Class-Driven Feature Selection
CDFS	Class Driven Feature Selection
CN	Condition Negative
CP	Condition Positive
EHR	Electronic Health Record
EM	Expectation Maximization
EMR	Electronic Medical Record
FN	False Negative
FNA	Fine Needle Aspirate
FP	False Positive
ICD	International Classification of Diseases
IG	Information Gain
IOM	Institute of Medicine
IWKNN	Improved Weighted K Nearest Neighbor
KNN	K Nearest Neighbor
LDA	Linear Discriminant Analysis
LVCF	Last Value Carried Forward
MAR	Missing At Random
MCAR	Missing Completely At Random
MI	Mean Imputation
MICE	Multiple Imputation by Chained Equations
MLE	Maximum Likelihood Estimation
MNAR	Missing Not At Random
MVDFS	Missing Variable Driven Feature Selection

PPV	Positive Predictive Value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PROBAST	Prediction model Risk Of Bias Assessment Tool
RF	Random Forest
ROC	Receiver Operating Characteristics
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
TS	Thoracic Surgery

Chapter 1: Introduction

This chapter includes six sections. The first section provides a background on healthcare data analytics, and how this field is evolving since the mandate of the us of the Electronic Health Record (EHR). The second section provides an overview of healthcare data analytics challenges. The third section provides an illustration of the research problem. The fourth section discusses the research scope and objectives. The fifth section explains the research significance and uniqueness. The last section provides an overview of the dissertation.

1.1 Healthcare Data Analytics

The field of big data and analytics is expanding rapidly in healthcare. The vast amounts of data collected in Electronic Medical Records (EMRs) contributes to and leads this growth, in addition to the improvement of the technical infrastructure and supporting computational power. This rapid development has enabled data scientists and analysts to extract new insights and data trends from the massive datasets collected and stored in healthcare databases. Data mining in the healthcare field aims to extract knowledge for effective medical diagnosis, prognosis, treatment, screening, monitoring and management (Jabbar, 2015). Healthcare data mining can be classified into two types: descriptive and predictive. Descriptive data mining includes clustering and association analysis, which allows for the exploration of hidden trends, groups, and relationships between features in collected datasets. Predictive data mining includes classification and regression models, which allows for modeling/prediction of outcomes or risks, provides an opportunity for healthcare workers to redesign processes to prevent the predicted outcomes, and supports management decision making.

Healthcare improvements driven by data mining have many examples, which include reducing patient harm (Dash et al., 2019), and improving patient experience and outcomes (Teo et al., 2021). The research of healthcare data analytics is growing and expanding. There is a recent growth of publications that integrate data science in healthcare. Figure 1.1 shows the number of publications in PubMed related to big data in healthcare in the past two decades.

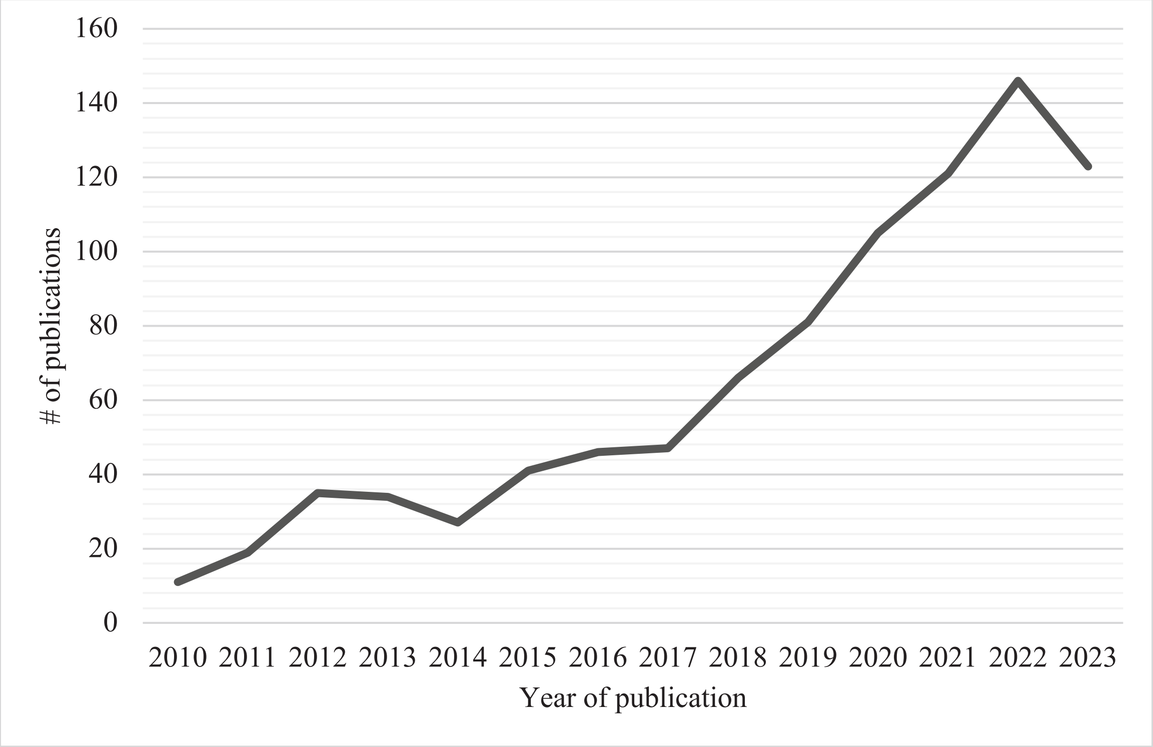


Figure 1.1: Number of PubMed Publications related to big data in healthcare (2010 - 2023)

Since the Electronic Health Record (EHR) mandate, the amount of data collected about patients, processes, and claims has increased. The large amounts of data have allowed for data science to be applied in this field in various ways. This includes descriptive, predictive, and prescriptive analytics. The integration of data science in healthcare is expected to revolutionize medical therapy and personalized medicine to improve healthcare outcomes and control costs (Dash et al., 2019).

The applications of data science in healthcare aim towards process improvement that include patient flow, appointment booking, and staff scheduling. Also, it aids in predicting patients' risk for adverse events and future diagnosis, in addition to forecasting patient volumes. Data science algorithms provide deep understanding of patient populations, through analyzing hidden trends in data to derive important and beneficial insights, improve hospital metrics, and support managers in decision making. It is expected that the implementation of data science in healthcare organizations could lead to 25% savings in healthcare organizations' annual costs (Dash et al., 2019). Better diagnosis and diseases prediction through predictive analytics can reduce hospitalizations and reduce cost. For example, a recent study proposed a readmission prediction model with estimated healthcare cost savings at over $1 million to prevent readmissions at a rate of 50% (Teo et al., 2021).

Healthcare data analytics include processing and utilizing all types of data, which include numeric, categorical, image, process metrics, molecular, and texts, that are related to patients and physicians. The data science algorithms applied on healthcare datasets are vast, they can be classified under data mining and machine learning. Data mining extracts knowledge from data using different algorithms. Machine learning focuses on the data learning process by computers and the development of algorithms that facilitate this learning. Deep learning is another method in data science that uses deep neural networks to process data to find trends and insights.

Machine learning methods are either supervised or unsupervised. Supervised learning algorithms can be used to discover the relationship between variables and one or more previously known outcomes. Supervised learning involves training a model using a training dataset that contains known outcomes. The goal is to train the model to classify or recognize outcomes using historical data. Unsupervised learning focuses on uncovering naturally occurring data trends or existing groups of data in an unlabeled dataset. Figure 1.2 illustrates three examples of machine learning uses in healthcare.

Section A of Figure 1.2 shows a supervised learning example to predict patient survival. This type of machine learning here is used to uncover the relationship between some clinical features and survivability using historical data. This is followed with a model designed to estimate survivability in future patients. Section B shows an example of unsupervised learning, which clusters similar data instances into groups. Deep learning in section C uses image data to extract and create features to represent information in a higher order of complexity to make predictions about those images, for example, to measure the probability of an existing tumor or abnormality in an X-ray.

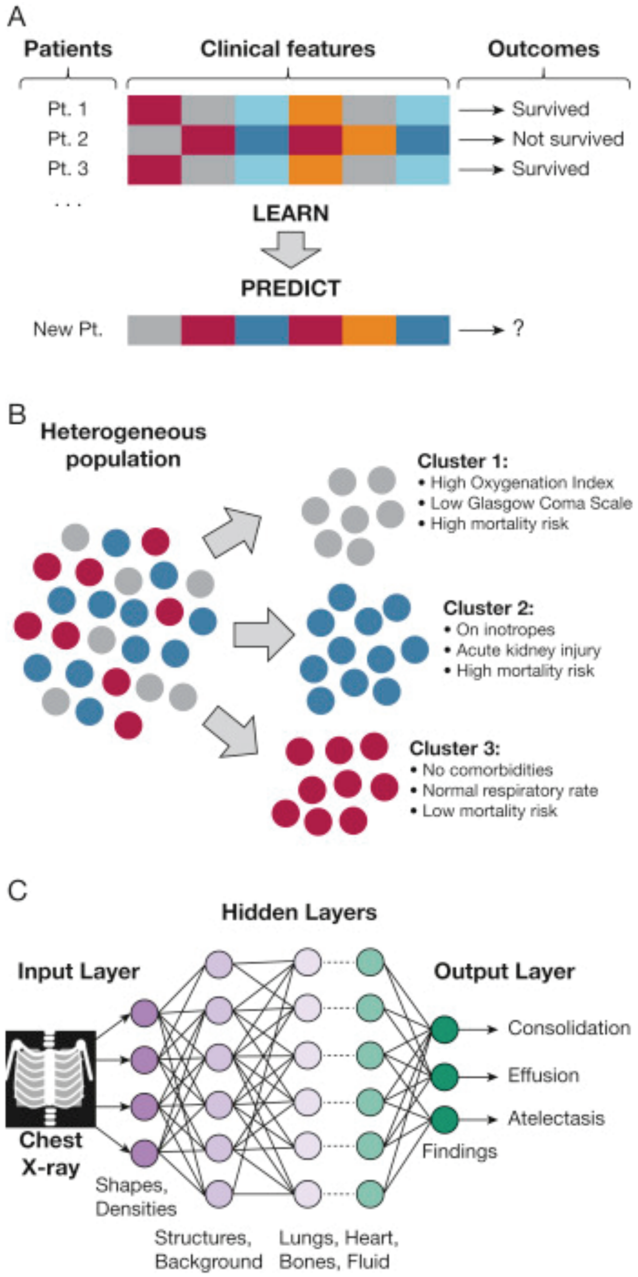


Figure 1.2: Examples of machine learning in healthcare settings (Sanchez-Pinto et al., 2018)

An overview of the steps of healthcare predictive analytics is provided in Figure 1.3. It starts with collecting vast amounts of data about a problem, which can include patient types, pathways, clinical or hospital scope. The data collection process involves data collection from multiple resources and systems where records can be linked together to provide a complete representative dataset of the subject of interest. After that, the collected data is preprocessed for handling errors, missing values, outliers, and incomplete records. The goal of preprocessing the dataset is to partially prepare the data for exploratory data analysis, data mining and machine learning algorithms. Some algorithms such as tree-based algorithms require the transformation of all the variables into categorical variables. Other algorithms such as logistic regression and K-Nearest Neighbor require scaling or normalizing numeric features to avoid bias towards features with higher values. After that, multiple algorithms are applied to the data to find the best performing model in terms of accuracy and model robustness. After selecting and tuning the best performing algorithm to build the model, a sample of data that was not used in training the model is used to measure the model's learning efficiency and predictability.

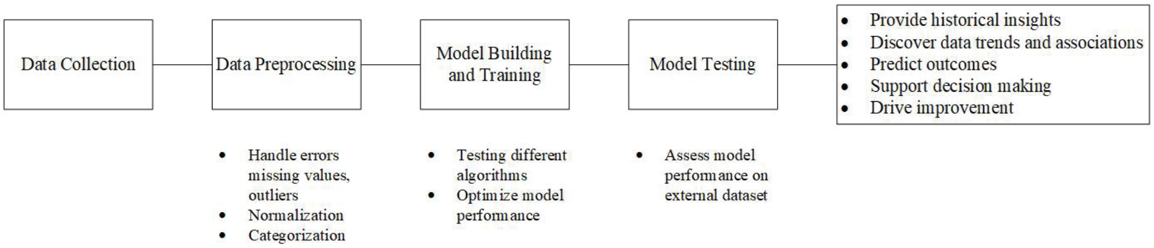

Data Collection
Data Preprocessing
Model Building and Training
Model Testing
• Provide historical insights
• Discover data trends and associations
• Predict outcomes
• Support decision making
• Drive improvement
• Handle errors missing values, outliers
• Normalization
• Categorization
• Testing different algorithms
• Optimize model performance
• Assess model performance on external dataset

Figure 1.3: Overview of steps of healthcare predictive analytics

The data collection and preprocessing steps in Figure 1.3 are often the most time and effort consuming steps in a predictive analytics framework. That involves multiple sub steps to ensure that the data collected is free of errors and shaped into the appropriate format before processing by data mining algorithms. Ferrao at al. (2016) proposed a roadmap for preprocessing clinical data for predictive modeling that was published in the *Applied Clinical Informatics Journal*, this roadmap is shown in Figure 1.4.

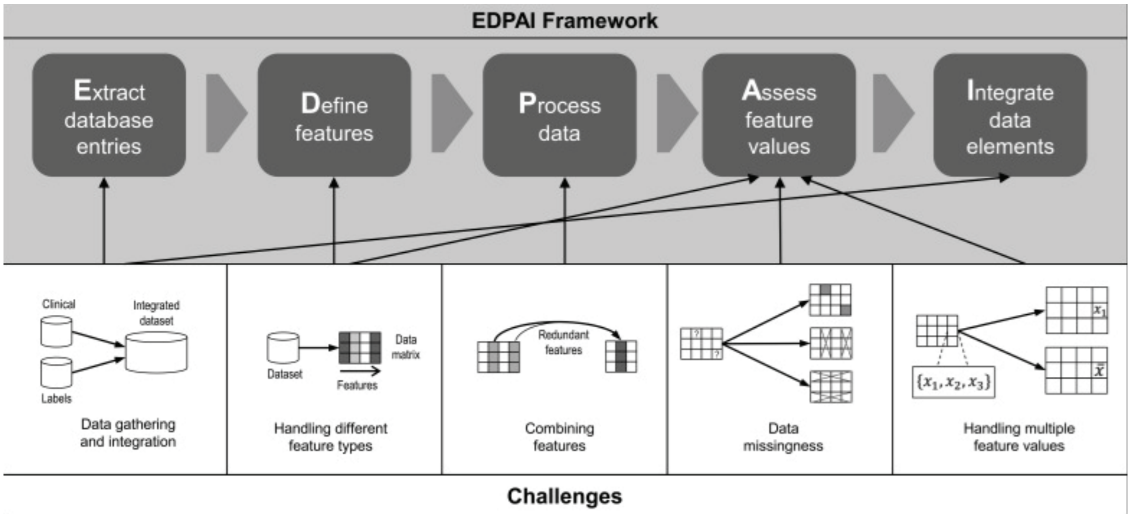


Figure 1.4: Steps of processing clinical data for predictive modeling (Ferrao et al., 2016)

The data extraction step can be time consuming because it requires gathering information from multiple resources and linking records together. Each variable definition should be clear to make sure it represents the required information; for example, chronic diabetes vs. new diabetes diagnosis are two different variables. After that, the collected variables go through processing steps, which include combining, calculating, and extracting features to make them readable by different algorithms, and to represent patients' data correctly and close to approximate reality. Handling missing data includes the investigation of missingness percentages and patterns between variables, as well as the imputation of missing values because they can introduce bias to models and can result in incorrect predictions. The last step is the integration of data elements, which puts the data in a final frame where each row represents an instance or patient, and each column represents a categorical or numerical label value.

1.2 Challenges of Healthcare Data Analytics

Applying data mining to healthcare data can be a challenging task (Jabbar, 2015). Healthcare data has different types of information that include medical variables, clinical variables, socio-economic factors, demographics, patient progress notes, images, and patient experience feedback. The majority of healthcare data is collected in the EHRs, which are the common source for extracting data for data mining applications. The challenges in working with healthcare data are about the high volume and imbalance of the data, as well as its unstructured format. Also, data is distributed among different systems, which can be challenging to connect. High missingness of variables and outlier percentage is a common problem in healthcare data. This is partly due to its proneness to human error and manual data entry, and the redundancy of collecting some variables with the lack of standardization (Duggal et al., 2016). The dynamic nature of healthcare data is one of the main challenges to data mining applications, where some variable values continuously change due to patient health conditions or health policies, and other variables are delayed and are not available until the patient is discharged.

Challenges facing healthcare data analytics were classified into three types: evidence-based, methodological, and clinical integration and utility challenges (Rumsfeld et al., 2016). Some of these challenges are intercorrelated and connected; overcoming one of them may help in overcoming another. These healthcare analytics challenges are illustrated in Figure 1.5.

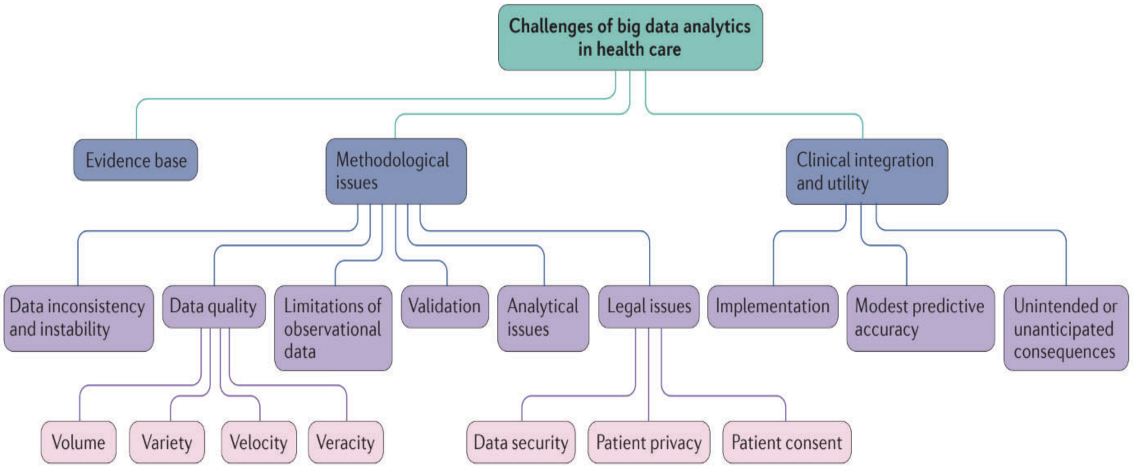

Challenges of big data analytics in health care
Evidence base
Methodological issues
Clinical integration and utility
Data inconsistency and instability
Data quality
Limitations of observational data
Validation
Analytical issues
Legal issues
Implementation
Modest predictive accuracy
Unintended or unanticipated consequences
Volume
Variety
Velocity
Veracity
Data security
Patient privacy
Patient consent

Figure 1.5: Challenges of predictive modeling in healthcare (Rumsfeld et al., 2016)

Healthcare data mining or data analytics publications have largely focused on the theories and concepts of building competitive predictive models with improved performance. There is a lack of evidence on the application or the application feasibility of these models (Rumsfeld et al., 2016), and how efficiently designed interventions drive improvement. Most of the research publications suggest that the use of these models would support the decision-making process and allow for early intervention and improve desired outcomes. The idea of identifying high-risk or high-cost patients could result in interventions that reduce patient risk, reduce costs, or improve outcomes cannot be assumed (Amarasingham et al., 2014). Also, this is highly connected to the validation challenge in Figure 1.5, in which the majority of published research internally validate proposed models through testing them on data sampled from initial collected data. In some studies, proposed models are also validated on external datasets to evaluate models' generalization measures. However, a very small number of models are actually implemented and validated through clinical workflow integration and demonstrate measurable improvement (Amarasingham et al., 2014).

The lack of evidence base problem and implementation under the clinical integration and utility section can be correlated. The lack of evidence-based on measurable improvements could reduce the interest in applying this field in healthcare to improve outcomes. The lower the number of implementation examples, the smaller that evidence-based piece becomes. It is a cyclic relationship where each one leads to the other, as shown in Figure 1.6.

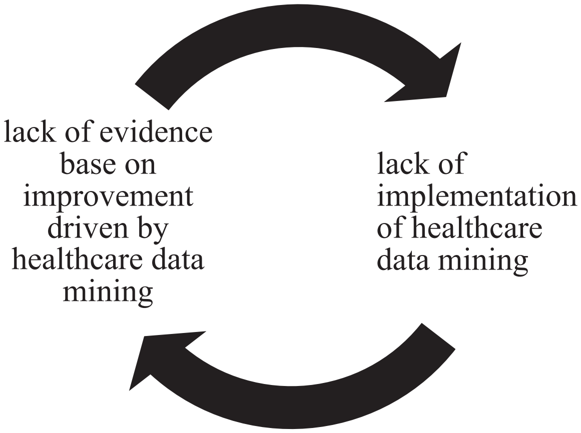

lack of evidence base on improvement driven by healthcare data mining
lack of implementation of healthcare data mining

Figure 1.6: Correlation between evidence-based improvement and implementation

The lack of implementation of healthcare data mining models in clinical settings can be due to multiple reasons that include the lack of evidence of models' effectiveness in improving healthcare outcomes and metrics. The clinicians' confidence and trust level in predictive models play an important role in the adoption and use of these models by healthcare providers. This can also be due to the lack of evidence and publications that discuss measurable improvement of healthcare metrics after utilizing similar predictive models. Moreover, implementing predictive models in healthcare faces other barriers associated with models' deployment and availability at the point of care even with the presence of an appropriate IT infrastructure and clinical support; these barriers are often due to missing or delayed data.

When data mining models are designed, historical data is collected from EMRs and other data sources to build a data repository that represents a certain scope of patients or a unit for a specific timeframe. The variables are preprocessed, which include handling errors, outliers, missing values and creating categories of variables, handling correlation, and others. Sometimes after that, important variables/features are selected through iterative processes to build a predictive model. Multiple iterations and types of models are tested on the dataset to find the best performing algorithm; this step uses a large subset of the collected data, in which approximately 20% to 40% of the data is usually kept for testing and validating the model's performance. Finally, knowledge can be discovered through the results of predictive models. Hidden data trends and associations between variables and target can be revealed.

The validation of the model in the abovementioned process is called retrospective validation, because the data is historical, in which all features are measured under the same static historical timeframe. Retrospective validation is a common approach to evaluate models' performance, which includes all types of internal validation such as cross validation, leave-one-out methods, and others. When a retrospective model is applied on prospective data, it may behave differently due to changes in model inputs overtime. Also, some variables, such as claims-based variables, could be missing or not available until a certain time (Amarasingham et al., 2014). This is the type of data missingness that acts as a barrier for deploying data mining models at an early point of care and for integrating them in the clinical workflow. Missing values are a common issue when analyzing data in a wide range of research fields. In the healthcare domain it is unavoidable. A significant amount of healthcare data is self-reported, where patients report their demographics and symptoms. In the manufacturing domain, most of the data is collected and documented automatically. In healthcare, the majority of the data is documented in patient charts by human entry, which makes the data prone to errors and loss.

There is a high value in deploying and integrating data mining models in the healthcare workflow. The availability of prediction scores at an early point of care provides an opportunity for intervention on high-risk populations and potentially prevent expected harm or adverse events. It provides healthcare providers with an opportunity to tailor individualized solutions for patients based on their unique case, which can prevent the predicted outcome. Also, the deployment of these models in a timely manner provides a prioritization criterion for caregivers when practicing common interventions. Both retrospective and early predictive approaches are beneficial for improving healthcare metrics, but validating a data mining model retrospectively is considered only the initial step in the journey toward the implementation of timely point-of-care prediction (Amarasingham et al., 2014). Figure 1.7 provides a comparison between the two approaches.

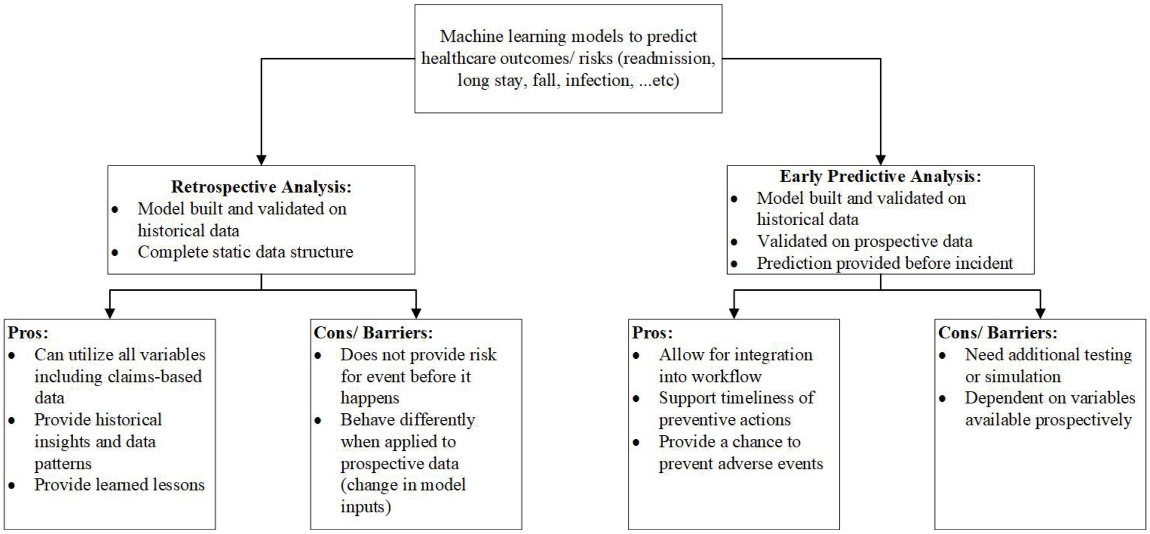

Machine learning models to predict healthcare outcomes/ risks (readmission, long stay, fall, infection, ...etc)
Retrospective Analysis:
• Model built and validated on historical data
• Complete static data structure
Early Predictive Analysis:
• Model built and validated on historical data
• Validated on prospective data
• Prediction provided before incident
Pros:
• Can utilize all variables including claims-based data
• Provide historical insights and data patterns
• Provide learned lessons
Cons/ Barriers:
• Does not provide risk for event before it happens
• Behave differently when applied to prospective data (change in model inputs)
Pros:
• Allow for integration into workflow
• Support timeliness of preventive actions
• Provide a chance to prevent adverse events
Cons/ Barriers:
• Need additional testing or simulation
• Dependent on variables available prospectively

Figure 1.7: Pros and cons for retrospective and early predictive healthcare models

If a retrospective model can be extended to an early predictive model, it would facilitate the timely utilization and integration of the model into doctors' workflow and routine. The added values would not be limited only to historical insights, they would also act as an early notification system about high-risk individuals, which allows for early intervention and potential prevention. The use of these models would contribute to the evidence of their applicability and efficiency in improving certain outcomes, which would increase healthcare providers' trust in these models and encourage them to use it, and thus potentially increase the evidence base as mentioned before.

1.3 Research Problem

As mentioned before, there is a lack of implementation of predictive models in healthcare in a timely manner to make predictions available at the point of care. Missing variables in prospective data act as a barrier for deploying predictive models, where models' inputs are not complete and thus cannot produce accurate predictions. Despite the fact that the literature is rich with predictive analytical models in healthcare, there is a limited number of publications that discuss the prospective validation and deployment of proposed predictive models in healthcare, and how models are successful in driving healthcare improvement.

After an extensive review of literature for the past 12 years (2010 – 2023), a limited number of research studies validated the effect of utilizing predictive models in real healthcare settings through timely deployment and workflow integration. Most of the publications designed predictive models retrospectively but rarely extended the research discussion to the usability and applicability of proposed models in healthcare settings. Of that, some extended the discussion to list barriers facing the deployment of proposed models and listed data missingness as one of the barriers. This will be discussed thoroughly in Chapter 2.

Most of the timely deployed predictive models in the literature focused on using early available data in patient records to predict outcomes. The studies mentioned in the literature provide evidence that there is an interest and a need for integrating predictive models in healthcare workflows in order to improve healthcare outcomes and metrics. Also, the deployed models show evidence that predictive analytics can drive measurable improvement when applied; however, more research is needed to increase this evidence base. Also, it is necessary to discuss the limitations for applying theoretical predictive models into operation and integrating them into workflow.

There is a need for solutions that can bridge this gap and move models from theory to deployment. The aforementioned points are indicators for a research gap and an opportunity that require further study and exploration. In a study that investigates external validation of clinical risk prediction models, missing data was listed as one of the main barriers for model implementation, where scores cannot be produced by models when some factors are not available or sparsely available (Hickey et al., 2016).

1.4 Research Scope and Objectives

Data missingness in prospective data is one of the main barriers addressed in the literature for integrating predictive models in workflow. This research focuses on tackling this problem and exploring methodologies to help solve for data missingness, which enables the deployment of predictive models at times that allow for preventive interventions. This research proposes a framework for handling missing variables in prospective datasets. It utilizes patients' historical information to provide a complete set of model inputs, which allows for deployment at an early stage of care episodes. The proposed approach includes comparison with common missing data handling methods, and it is applied on multiple datasets to insure robustness and generalizability. The datasets selected have different levels of dimensionality, number of instances, number of features, and class balance. Also, the proposed approach will be tested and applied on different levels of missingness, i.e., different numbers of missing features, which reflect robustness.

This research aims to answer the following questions:

1. What are the impacts of missing prospective data on predictive models' deployment and performance?
2. How can complete retrospective datasets be utilized to impute missing variables in prospective datasets? How would that compare to standard missing data handling approaches?

3. How far are the values generated by imputation methods from original values? How does this change across imputation methods?
4. How does the number of imputed variables impact models' performance? Does that vary by imputation method?

The objectives of this research are:

1. Design a data mining framework to handle missing variables in prospective datasets for classification problems;
2. Integrate feature importance with K-Nearest Neighbor (KNN) for imputing missing variables;
    a. Feature importance in terms of classification target/class; top features that contribute to prediction.
    b. Features associated with prospectively missing features.
3. Demonstrate the generalizability of the proposed approach on datasets with varying characteristics that include number of instances, number of features, and class imbalance.

1.5 Research Significance and Uniqueness

The significance and uniqueness of this research comes from its facilitation that allows variables from prospective data to be utilized in designing predictive models even when delayed. The Institute of Medicine's (IOM) six aims of healthcare quality include safety, effectiveness, patient-centeredness, timeliness, efficiency, and equity. This research aims to improve the timeliness and efficiency aims through solving one of the major barriers for providing timely predictive analytics in healthcare; this would assess in delivering and improving care in a timely efficient manner, increase the chances of preventing patient harm, and reduce costs and wastes in healthcare.

This research proposes a hybrid data mining framework that integrates feature selection with KNN to impute missing variables in prospective datasets. The proposed framework includes two approaches that modify the search space of the KNN imputation algorithm to find similar observations. The first approach utilizes wrapper feature selection. The second approach utilizes filter feature selection.

The proposed framework reduces the deterioration in classification models' performance in prospective datasets with up to 5 missing variables imputed, compared to common imputation methods such as MI and KNN imputation. It maintains classification models' performance closer to internal validation as the number of missing prospective variables increases.

This research ensures robustness and generalization of the proposed approach through testing it on different levels of missingness in datasets with varying characteristics. This includes varying number of instances, number of features, and class imbalance levels.

This approach aims for bridging the gap between theoretical retrospective models and early preventive models in healthcare. This brings new opportunities to different healthcare areas for driving continuous improvement initiatives through predictive analytics workflow integration.

1.6 Dissertation Overview

This chapter provides an overview of healthcare data analytics and common challenges in this field. It explains the problem statement of this research and its objectives. Chapter 2 provides a systematic literature review of publications since 2010. Chapter 3 explains the research framework followed to for handling prospectively missing features. It provides a detailed explanation of the steps followed to simulate prospectively missing features in multiple datasets. It illustrates the steps for imputing missing data using two novel approaches proposed in this research, and two existing data imputation methods. Chapter 4 includes the results of applying imputation methods on multiple datasets and provides a comparison between all methods. Chapter 5 summarizes the conclusions and future work of this research.

Chapter 2: Literature Review

This chapter discusses various machine learning and data mining research that is conducted in the field of handling missing data in healthcare, which includes a focus on prospective missing data in EHRs. The literature review in this dissertation follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA).

2.1 Missing Data in Healthcare

Utilizing the massive amounts of data collected in healthcare in EHRs can be challenging due to the incomplete nature of patient records. The reasons of missing values in healthcare data are tied to many factors, which include the lack of standardization in documentation, timing, and sequence-related factors. Also, it is prone to human error because much HER data is inserted manually. The rate of missing data in EHRs was reported to be between 20% and 80% (Hu et al., 2017). This affects the conclusions and results obtained from data analysis and analytics, where high data missingness significantly affects results' robustness and generalization and introduces bias. Some researchers categorize missing data mechanisms into three main types based on the missingness characteristic: Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not at Random (MNAR) (Hu et al., 2017). Table 2.1 explains each type in detail.

Table 2.1: Types of data missingness

Type	Definition	Example
MCAR	The cause of missingness is not associated with any data characteristics or variable values; it is randomly missing	Missing test result due to randomly broken blood sample

MAR	The cause of missingness is associated with the value of another variable.	Missing data for males in a patient satisfaction survey, but it does not have an impact on their satisfaction level
MNAR	The cause of missingness is associated with the value of the missing value itself	Patients with disabilities are more likely to refuse to complete patient experience survey

This chapter provides a detailed review of the previous research for the past 12 years (2010-2023) that focused on handling missing data in EHR that is utilized in machine learning and data mining, and how data missingness impacted the ability to deploy machine learning models prospectively in healthcare environments.

There are multiple methods used in the literature to handle missing data in healthcare studies. Some methods are focused on imputing or predicting missing values using statistical methods or machine learning algorithms. Other approaches include deleting records with one or missing value. these methods are summarized in Figure 2.1.

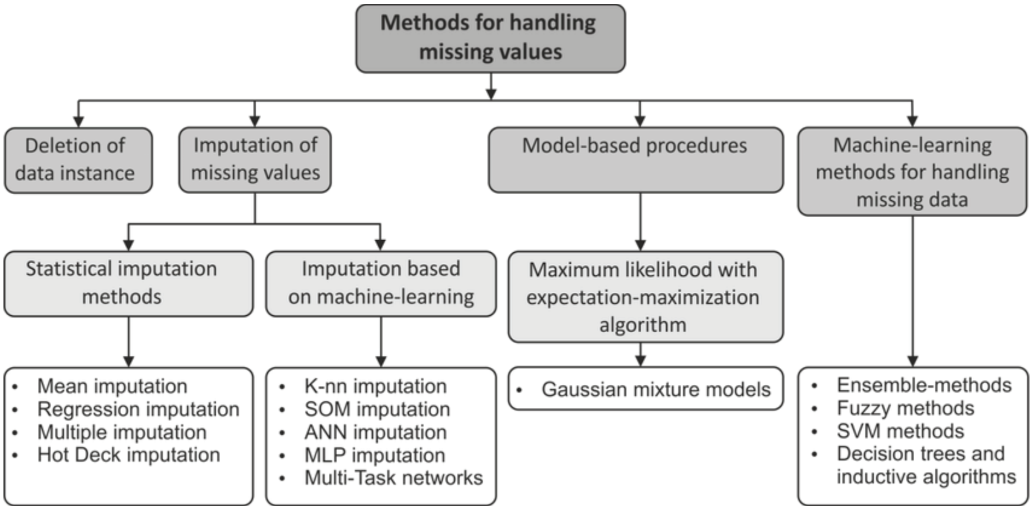

Methods for handling missing values
Deletion of data instance
Imputation of missing values
Model-based procedures
Machine-learning methods for handling missing data
Statistical imputation methods
Imputation based on machine-learning
Maximum likelihood with expectation-maximization algorithm
• Mean imputation
• Regression imputation
• Multiple imputation
• Hot Deck imputation
• K-nn imputation
• SOM imputation
• ANN imputation
• MLP imputation
• Multi-Task networks
• Gaussian mixture models
• Ensemble-methods
• Fuzzy methods
• SVM methods
• Decision trees and inductive algorithms

Figure 2.1: Methods of handling missing data (Garcia et al., 2010)

Deletion or complete-case analysis is one of the commonly applied methods in healthcare studies, in which observations with at least one missing value are excluded from the study. This method can result in reducing the size of the dataset and may lead to biased results, inaccurate conclusions, and loss of power.

Imputation of missing values includes two types: statistical imputation and machine learning imputation. Simple statistical imputation is replacing the missing value with a calculated value from the dataset, such as the mean of complete observations (Mean imputation), or through fitting a regression model to estimate the missing value (Regression imputation). Statistical multiple imputation accounts for the uncertainty and variable ranges that a true value can take through multiple iterations of imputation. Multiple Imputation by Chained Equations (MICE) is one of the primary multiple imputation methods. MICE is an iterative method that repeatedly applies simple mean imputation on portions of the dataset. The values that best fit the data distribution are selected. This method maintains a similar data distribution before and after imputation. However, categorical variables must be converted to dummy variables and then rounded to closest binary number.

Machine learning imputation includes utilizing a predictive algorithm to predict the missing value and impute it. For example, KNN is a method that searches for similar near observations using Euclidean distance or other distance measures, then imputes the missing value with the mean or mode of the nearest neighbors. The number of neighbors and weights are predefined. Categorical variables must be transformed to dummy variables, and data must be normalized to avoid bias.

Model-based procedures such as the Gaussian mixture models is an algorithm that starts with estimating the dataset distribution using complete instances. Then missing value estimation is done accordingly. Maximum Likelihood Estimation (MLE) is a method that estimates the parameter values of a model in a way that increases the probability of obtaining observed data. This method is used to estimate model parameters using complete observations (Full Information Maximum Likelihood) and then used to borrow information from complete observations to incomplete ones while maintaining the accuracy and precision of estimations (Enders et al., 2001). Expectation-Maximization (EM) is another method that performs iterations to find the maximum likelihood estimates of model parameters. EM and MLE are very similar approaches in that they both aim for finding the best estimation of model parameters, but EM starts with a random selection of parameters while MLE uses the observed dataset.

Machine learning methods include the use of a predictive model or an ensemble model to predict the value of the missing value. Additionally, there are other methods such as the Last Value Carried Forward (LVCF) method. It is one of the common approaches that impute the most recent available value of a variable if the value was missing. Other methods keep missing values in the dataset. This allows models to learn from data missingness patterns; however, this method does not apply to all data mining and machine learning algorithms.

2.2 Handling Missing Data in Healthcare Systematic Review

A review of the literature published between 2010 and 2023 was conducted following Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines (Moher et al., 2009). This review covered journal articles and conference proceedings in searchable databases that include PubMed, ScienceDirect, and Web of Science to get a comprehensive systematic review. The research used the following keywords: "missing data", "missing values", "incomplete data", "healthcare", and "health care".

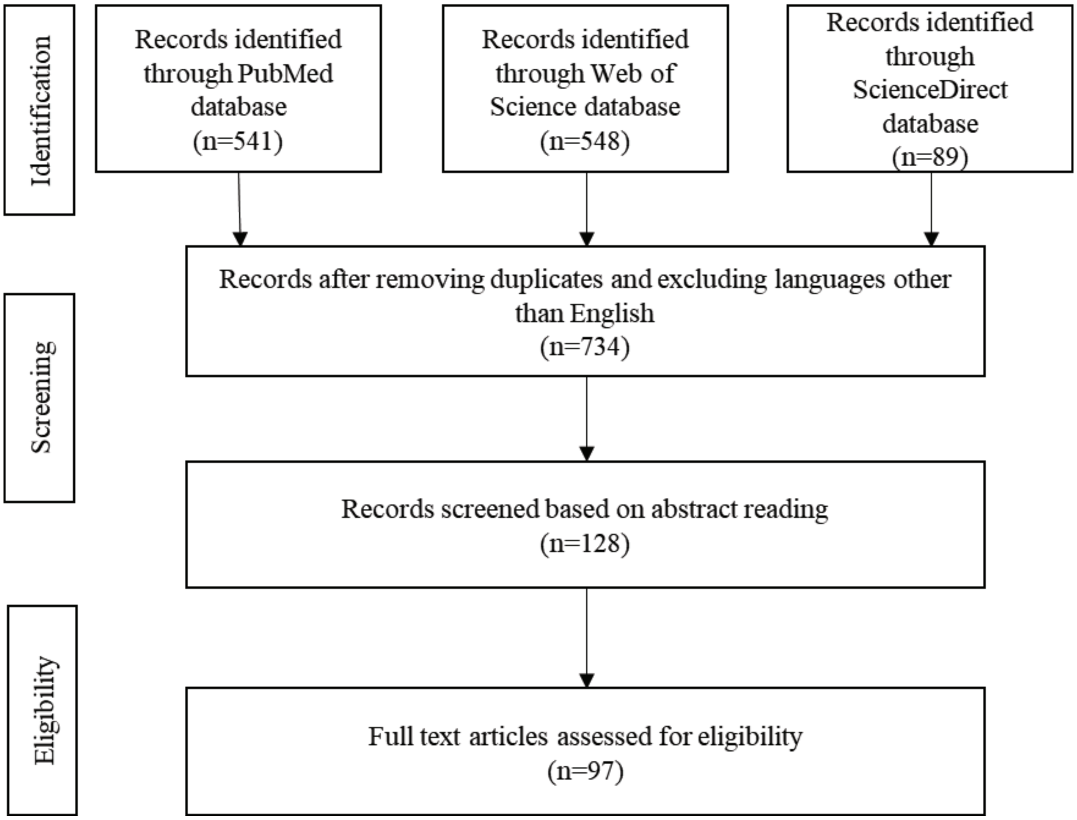

Identification
Records identified through PubMed database (n=541)
Records identified through Web of Science database (n=548)
Records identified through ScienceDirect database (n=89)
Screening
Records after removing duplicates and excluding languages other than English (n=734)
Records screened based on abstract reading (n=128)
Eligibility
Full text articles assessed for eligibility (n=97)

Figure 2.2: Literature search framework per PRISMA guidelines

The term predictive analytics was not included in the initial keyword search because there are many synonyms for this concept including big data, data science, data mining, machine learning, and more, therefor the scope was narrowed down by screening and reading abstracts and contents of publications individually. The initial search of the three databases resulted in a total of 1,178 publications. Figure 2.2 shows the steps of the literature search based on PRISMA guidelines. Papers that were written in languages other than English were excluded, and the articles without access to the full text were also excluded. After screening abstracts, papers that were out of the scope of this research were excluded, the final number of publications in scope was 97. The 97 articles included three systematic literature review publications that focused on handling missing data in healthcare in general; these are discussed in section 2.2.1.

Literature survey publications were reviewed to learn about the identified challenges associated with missing data, and the methods used to overcome these challenges in healthcare predictive analytics. Also, reviewing literature survey publications provides details about the evaluation criterion authors used to evaluate reviewed publications. The other 94 publications were research studies that focused on healthcare predictive analytics and missing data handling. Figure 2.3 explains the classification of the publications in scope.

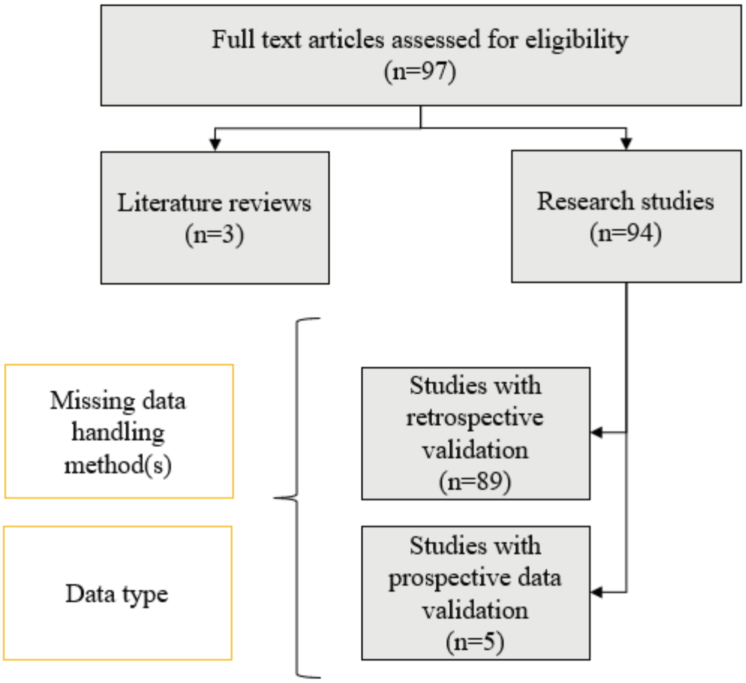

Full text articles assessed for eligibility (n=97)
Literature reviews (n=3)
Research studies (n=94)
Missing data handling method(s)
Data type
Studies with retrospective validation (n=89)
Studies with prospective data validation (n=5)

Figure 2.3: Classification of articles in scope

Studies were classified into two categories: literature reviews (n=3) and research studies (n=94). The research studies were broken down into two categories: studies that applied retrospective validation (n=89) and studies that prospectively validated their models (n=5). All research studies were classified into categories based on missing data handling methods and data type. The 94 studies will be discussed thoroughly in section 2.2.2

2.2.1 Previous Literature Surveys on Missing Data in Healthcare

After extensive research in the literature for the past 13 years, three articles were found that conducted systematic reviews for handling missing data in healthcare in general. These studies are summarized in Table 2.2.

Table 2.2: Literature survey publications

Citation	Review Scope	Review Years	Key Findings
Tanna et al., 2020	Heart failure risk prediction models	2013-2018	• Used PROBAST to evaluate models (40 studies, 58 models) • Few reported details on handling missing data (n=11; 28%) • 81% of models had lacked areas that caused high risk of bias due to insufficient description of missing data handling, validation, and calibration
Chee et al., 2021	COVID-19 AI applications in intensive care and ED	2020	• Used PROBAST to evaluate models (14 studies, 11 models) • 11 predictive modeling studies; **all had high risk of bias** • **Poor missing data handling** and **weak model validation** contributed significantly to bias • None have been validated in ED or intensive case settings • Current COVID AI applications are **not ready for deployment**, need effective clinical adoption

Nijman et al., 2022	Studies developing/ validating clinical ML prediction models	2018 - 2019	• 152 included studies; **37% did not report missing** data details • Most common approach was **deletion** (comp. case analysis); likely **caused bias and loss of predictive power** • The majority **did not** report **sufficient information on missing data presence or handling** • Prediction model researchers **should be more aware of alternative methodologies to address missing data**

The first publication conducted a systematic literature review for evaluating heart failure risk prediction models published between 2013 and 2018 (Tanna et al., 2020). The authors of this research considered missing data handling and the extent of models' validation for evaluating these prediction models using the Prediction model Risk of Bias Assessment Tool (PROBAST) (Wolff et al., 2019). The results show that only 28% (11 out of 40) of the reviewed models discussed how missing data were treated, and 55% (6 out of 11) of those applied multiple imputation. According to the application of PROBAST, 81% of reviewed models were not sufficiently described, which makes the quality and robustness of these models questionable; the identified lacking areas included model validation and missing data handling and the two are interrelated. The authors confirmed that there is a real need to integrate risk prediction models in healthcare workflow and missing values form a barrier to these models.

In 2021 a systematic literature review study reviewed and evaluated COVID-19 artificial intelligence applications in intensive care units and emergency settings published in 2020 (Chee et. Al.,2021). Eleven predictive models were in the scope of this study. All of them were evaluated using the PROBAST evaluation tool. It was found that all 11 models were rated at high risk of bias and mainly for poor missing data handling and weak model validation. Also, the authors stressed the fact that these models were not validated in ED or intensive care settings and thus were not ready for deployment.

Studies that developed or validated clinical machine learning prediction models for the years 2018 and 2019 were systematically reviewed (Nijman et al., 2022). It was found that 37% of the reviewed studies (n=152) did not report anything on missing data, and for the ones that mentioned missing data handling techniques, the most used approach was deletion, which likely causes bias and loss of predictive power. It was ruled that the majority did not report sufficient information on missing data presence or handling. The authors advised that researchers should be more aware of alternative methods to address missing data. This would allow readers to develop a better understanding of how these models were built and evaluate the feasibility of these models to be deployed in healthcare floors.

There is limited research content that focused on the healthcare missing data problem and there is a limited number of articles that systematically reviewed methods for handling missing data in healthcare predictive analytics in general. This fact motivated this dissertation to conduct a systematic review of the literature to investigate the missing data problem in healthcare from different angles. This would help in identifying the common methods used to handle missing data in healthcare predictive analytics applications. This would also help investigate how often missing data in prospective datasets is addressed for models' applicability in healthcare environments. Additionally, review the impact of prospective missing data on healthcare predictive models’ performance.

2.2.2 Healthcare Predictive Analytics and Missing Data

Figure 2.4 shows the distribution of all research studies (n=97) between the years 2011–2023 with a trendline. The number of publications follows an increasing trend. Table A in the Appendix provides details about the methods used for handling missing values and data types for the studies that are not covered in detail in this chapter.

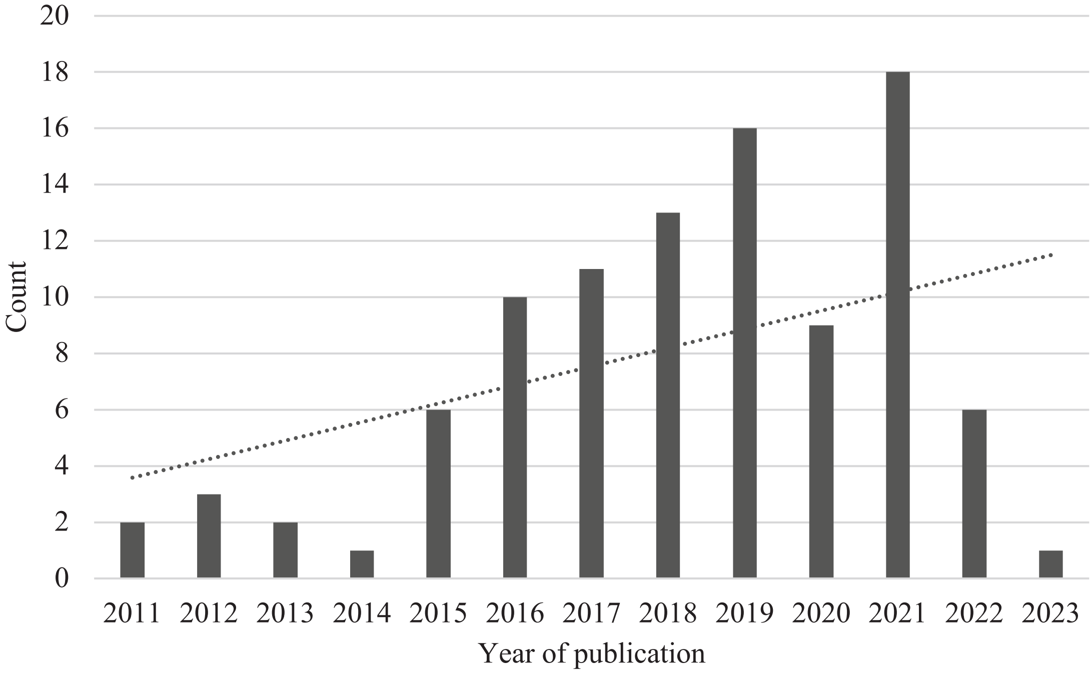


Figure 2.4: Number of reviewed publications distributed among years 2011–2023

The 94 research studies provided details about handling missing data as a part of their healthcare predictive analytics. Some studies applied and compared among many missing data handling methods. For that reason, each method application was counted. There were 168 applications of a method to handle missing data in the 94 publications. Table 2.3 shows each method and the count of publications and percentage that used the method.

Table 2.3: Number of times a method was applied in one or more of the publications in scope.

Method	Count of publications that applied method	% of applied methods (170)
Machine Learning Imputation	44	26%
Statistical Multiple Imputation	30	18%
Complete Case Analysis/ Deletion	30	18%
Statistical Single Imputation	25	15%
Other Methods	13	8%
MLE/ EM	11	6%
Keep missing values	12	7%
LVCF	5	3%
Grand Total	170	100%

Machine learning imputation was applied by 44 studies out of 94. It was the most used method followed by multiple imputation, deletion, and statistical single imputation, respectively. The focus of this research is imputation, statistical and machine learning-based, in EMR data.

In total, there were 63 publications out of 94 that applied statistical or machine learning imputation methods to handle missing data. 50 out of the 63 was in data obtained from EHRs. Table 2.4 and Table 2.5 explain the breakout of these articles, respectively.

Table 2.4: Count of research articles that applied listed missing data handling methods

Methodology	Number of publications
Imputation	63
Deletion	29
Other methods	34

Table 2.5: Classification of imputation research studies among types of data

Data Type	Number of publications
Numerical, Categorical (EHR)	50
Time series (Mobile Health Devices, IoT)	9
Other (simulated dataset, survey, etc.)	6

In Table 2.4, the classification of missing data handling methods was inspired by the classification of methods provided by (Garcia et al., 2010) as illustrated previously in Figure 2.1. Imputation includes statistical imputation and imputation based on machine learning. Deletion includes complete case analysis. Other methods include model-based procedures, machine learning methods, LVCF, keeping missing values in the data, and other. Details about methods used for each study is provided in the Appendix.

It is worth mentioning that most studies applied or tested imputation methods on training sets (or internal validation) that were collected retrospectively. Only three studies out of 50 tested the imputation on a prospective dataset, or simulated missingness in a prospective sample and applied imputation. These studies will be discussed in section 2.2.3. Table 2.6 provides a list of key publications that applied data imputation on EMR data retrospectively. The table includes studies' objective, imputation method, results, and gaps.

Table 2.6: Key publications that applied imputation in EMR predictive models

Citation	Objective	Method	Key Results	Gaps
Yang et al., 2016	Predict no-reflow after cardiac surgery among multiple types of variables with missing values (numeric and categorical)	Improved Weighted KNN (IWKNN) imputation (mean and mode of KNN)	Effective model for no-reflow compared to LR, LASSO, Artificial Neural Networks	• Variables with 20% missingness were dropped. • No details about handling missing values for other prediction methods
Chowdhury et al., 2017	Present a general model for the imputation of all types of missing data	Amelia (bootstrapping and Exp-Max), FURIA (fuzzy if-then rules), MICE applied on the gender attribute in three real datasets	MICE outperformed other methods	• Simulated/ tested missingness in one variable only
Devin et al., 2018	Identify clinical factors associated with return to work at three months among cervical spine surgery patients	MICE + LR used to identify factors associated with higher odds of returning to work	High AUC value for LR model	• Internal validation • Only one imputation method applied
Stiglic et al., 2019	Build a prediction model for type 2 diabetes with existing missing data	LASSO Regression + imputation using MissForest or Deletion at varying missing data rates	Major deterioration in model performance as missing rates increase	• Simulated missing rates across whole dataset
Garies et al., 2020	Test methods for improving the quality of patient smoking data in EHR data	MICE + pattern matching algorithm. External validation using population survey (comparing proportions of	Imputed data and survey data have similar categories	• Significant remaining missing data not treated (23.6%) • Approach tested in one variable

		smoking categories with general population survey)		
Jian et al., 2021	Predict eight diabetes complications (eight binary classifiers)	Deletion (Variables: ≥ 40% missing. Records ≥ 60% missing). Mode Imputation for categorical. For Numerical vars: MI, KNN, MissForest	MissForest was selected due to lowest imputation RMSE.	• Simulated missing data at same missing rate for incomplete records (4.4%) • No prospective application/ simulation.
Guo et al., 2021	Mortality prediction in cirrhosis patients	Prediction: DNN, RF, LR. Imputation: MICE and mode for categorical variables	Increased accuracy in predicting mortality for DNN compared to MELD*	No prospective application/ simulation
Pang et al., 2021	Predict early childhood obesity using ML	Prediction: XGBoost (selected out of seven algorithms). Imputation: keep, LVCF, MICE, **KNN**	MICE changed features' distribution and generated clinically implausible values. KNN reserved distribution	• No prospective application/ simulation
Payrovnaziri et al., 2021	Evaluate the impact of missing values imputation methods in EHR data on predictive models' performance and interpretation	Imputation: MI, mode, KNN, MissForest, DNN.	A small change in RMSE does not mean small change in model performance. Imputation methods should not be selected based on best accuracy only, but also robustness across higher missing rates	• Random missing data simulation across all features • No prospective application/ simulation
Nijman et al., 2022	Evaluate imputation and regression	Imputation: MI, KNN, and MICE. Fit 3	Can handle large amounts of missing	• Feature-wise missing rate in

	methods for training and testing data	imputations against original value (XGB Regressor). Feed three imputations of test set to XGB model to select best prediction (lowest RMSE)	data. Ensemble model outperformed single methods	testing set reached up to only 50%

The evaluation of missing data imputation in the reviewed studies was done in multiple ways. Some studies introduced different levels of random missingness in complete datasets, followed by measuring similarity (error) between imputed and original observations. Others applied multiple methods for imputing missing values, then compared the performance of the predictive model across imputation methods. Some tested combinations of machine learning models and data imputation techniques. Then, they selected the combination that resulted in the best classification accuracy. Lastly, some used a single method to impute missing data without comparing it to other methods.

2.2.3 Prospective Predictive Analytics and Missing Data

Based on the reviewed publications, only three publications out of 94 discussed the deployment of predictive models proposed with handling missing EHR data. Table 2.8 summarizes these publications.

Table 2.7: Publications that deployed or discussed deployment plans of predictive models with the presence of missing data

Publication	Objective	Imputation Method	Results	Gaps/ Limitations
Groenhof et al., 2019	Cardiovascular risk prediction	Mean imputation using training set mean	Model embedded in EHR system. Failed on deployment due to missing data	No other imputation methods were tested, variables not available on deployment.
Cesario et al., 2021	Early sepsis risk prediction	LVCF + Mean imputation	AUC=91% and increased to 97% on third day of stay	Limited to variables available on admission
Stock et al., 2021	Predict preterm birth using medical test reading + risk factors	Multiple Imputation (MICE)	Improved AUC compared to medical test by 5%	Limited to variables available on admission

The first study in Table 2.8 was able to design and deploy a cardiovascular risk prediction model in a real-time manner (Groenhof et al., 2019). The model was integrated with the EHR system to provide early predictions integrated with the clinical workflow. Missing data on deployment was handled using mean imputation. The calculation of risk estimates and treatment suggestions failed. The reasons for data missingness were delays or data that resides in temporary locations as unstructured data. The authors described the prospective missing data issue as a limitation to using risk prediction algorithms. They addressed the need for exploring methods to handle prospective missing data. They also mentioned that it is important to discuss with physicians who use the models the missing variables and applied imputation approaches.

The second study included building a prediction model to predict the risk for sepsis early in the care episode (Cesario et al., 2021). Variables were available early on admission, which allowed for building the model retrospectively and deploying it for active patient records.

A similar approach was used in the second study where variables from the EHR were combined with the concentration of vaginal fluid fetal fibronectin (quantitative fFN) medical test value to predict risk for preterm birth in women with preterm labor symptoms (Stock et al., 2021).

Using variables available on admission or early in the care episode is one way to avoid dealing with missing data on deployment. However, it could impact the performance reached for predictive models as some variables can increase the predictive power of the predictive models. Also, researchers and data scientists are not always able to produce accurate robust predictive models using a limited set of variables constrained by availability on admission or deployment. Models should be built with all possible informative variables included, and missing variables on deployment is an issue that requires further investigation and exploration.

Missing data affects the performance of predictive models in multiple ways; it introduces bias into model predictions and can reduce its classification accuracy. Although there are methods to handle missingness through deletion or imputation, these methods affect the performance of predictive models (Yadav et al., 2018).

In 2021 a study was conducted to analyze the impact of missing data imputation on predictive models' performance (Payrovnaziri et al., 2021). Many data imputation methods were applied on a complete dataset with simulated missingness to investigate how imputation changes data values compared to original values, and how the new values change model performance compared to original values. The experiments of this study revealed that a small change in RMSE for imputation is not always associated with a small change in model performance metrics. Imputation methods highly impact model performance metrics. However, these methods should not only be selected based on accuracy, but also based on robustness across higher rates of missing data.

There is a research demand for conducting studies related to predictive analytics with the presence of missing data. Three research areas in need of extended exploration were identified, which are related to the performance of predictive models with missing data (Nugroho et al., 2019):

1. Complete historical dataset, new data with missing values
2. Historical dataset with missing values, complete new dataset
3. Historical and new datasets with missing values

Missing data can limit the extension of a predictive model into deployment and integration in workflow, mainly in healthcare. Healthcare predictive models are trained on historical datasets that represent patient records, hospital stays, or appointments. These historical datasets are collected under a historical timeframe to ensure maximum data completion rates (lower missingness rate) such as complete care episodes or accomplished patient visits. When models are built successfully and predict target variables accurately, the ideal option would be to move models into operation, which provides early predictions and drive improvement, interventions, and preventive actions. However, in reality, to apply these models to prospective data, some variables could be missing and simply cannot, or cannot yet exist. That would impact the performance of prediction models significantly. This scenario can be linked to the research areas 1 or 3 discussed previously (Nugroho et al., 2019).

2.4 Summary of Related Literature

Through reviewing the research related to handling missing data in healthcare predictive models, it was noticed that a very limited number of publications extended their research to tackle models' deployment in healthcare settings, either by proposing deployment plans or testing models on real-time data. In the publications that discussed deployment, data missingness was identified as a major limitation holding predictive models from deployment and can affect their performance and adoption.

It is important to highlight differences and reasons behind data missingness in historical and new datasets, because missingness in historical data fields can be related to reasons such as technical systems' downtime, or lack of a need for certain variable collection (for example lab tests not needed). However, in new datasets or data at model deployment, the presence of missing values can be related to additional reasons such as delays, data stored temporarily in an unstructured format, which includes progress notes. Also, it can be dependent on the International Classification of Diseases (ICD-10) coding that would not be available until patients are discharged. The missingness of EMR data on deployment should be taken into consideration when predictive models are designed to drive interventions and improve care outcomes.

Chapter 3: Methodology

This chapter describes the methodology and results of the proposed approaches that combine feature selection with the KNN imputation algorithm. It starts with explaining the research methodology framework. Then it covers details about feature importance, feature selection, and the different types of feature selection algorithms. Additionally, it provides introduction about missing data imputation methods with a focus on the KNN imputation method.

3.1 Research Framework

This research proposes an integrated method that combines feature selection techniques with KNN imputation to handle prospective missing data. The proposed methodology maintains the performance of classification models when applied on prospective datasets with missing features. Figure 3.1 illustrates the research framework, which consists of four main steps: data extraction and preprocessing, feature importance calculation, data missingness simulation and imputation, and evaluation and assessment.

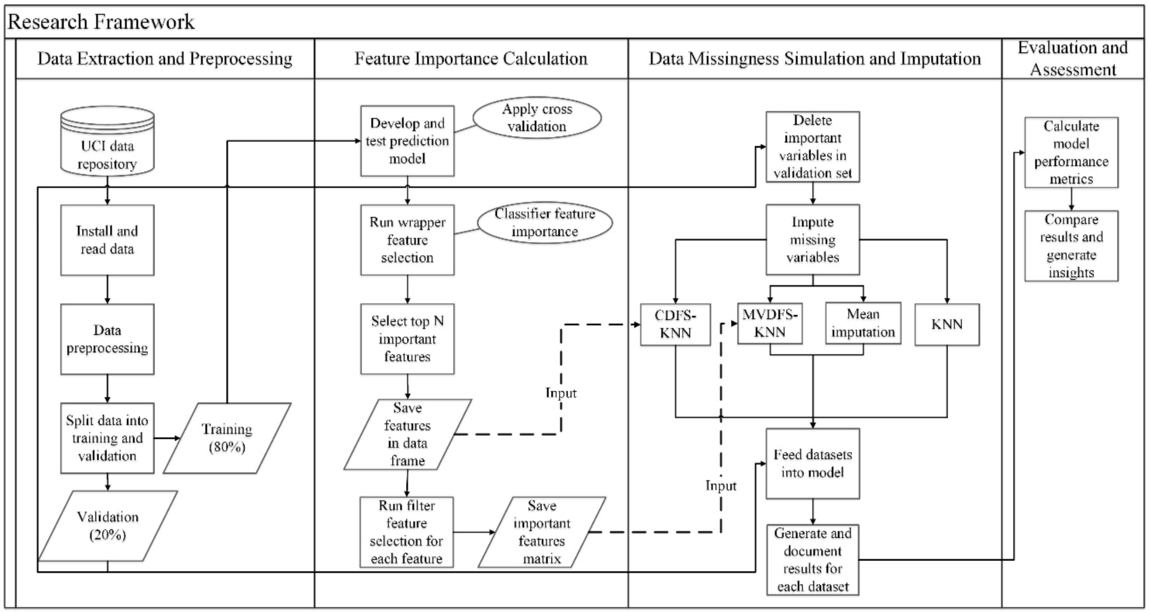

Research Framework
Data Extraction and Preprocessing
Feature Importance Calculation
Data Missingness Simulation and Imputation
Evaluation and Assessment
UCI data repository
Install and read data
Data preprocessing
Split data into training and validation
Training (80%)
Validation (20%)
Develop and test prediction model
Apply cross validation
Run wrapper feature selection
Classifier feature importance
Select top N important features
Save features in data frame
Run filter feature selection for each feature
Save important features matrix
Input
Input
Delete important variables in validation set
Impute missing variables
CDFS-KNN
MVDFS-KNN
Mean imputation
KNN
Feed datasets into model
Generate and document results for each dataset
Calculate model performance metrics
Compare results and generate insights

Figure 3.1: Research framework

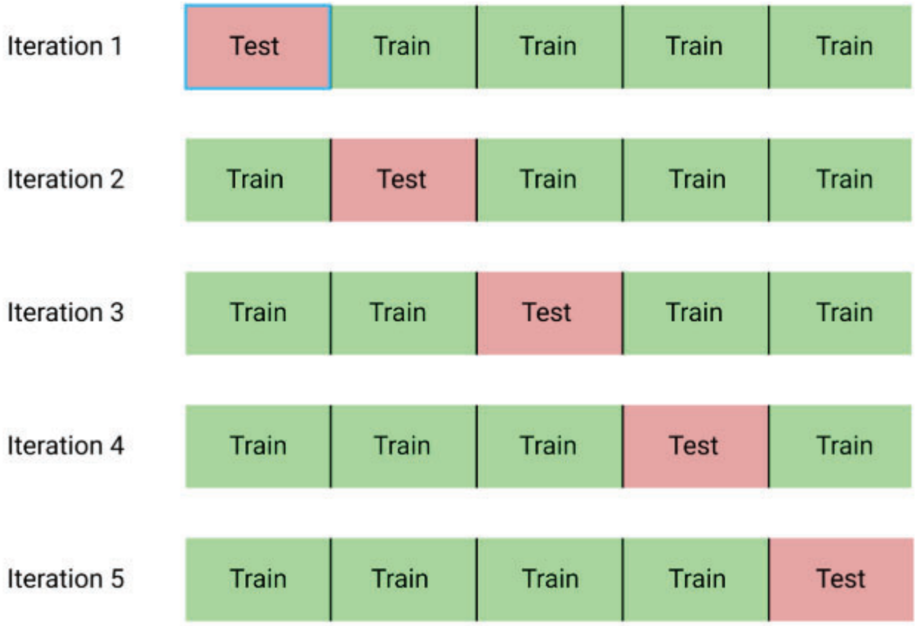
The first fold of this dissertation focused on imputing prospective missing variables through utilizing the KNN imputation algorithm with a modified search space. Wrapper feature selection was applied to modify the KNN search space of nearest neighbors. This approach is called Class-Driven Feature Selection KNN Imputation (CDFS-KNN). The performance of the CDFS-KNN imputation was compared to the traditional KNN imputation and mean imputation because these methods are some of the commonly used methods in the literature. This approach was tested on varying levels of missingness, starting with one missing variable, and up to five missing variables.

The second fold of this dissertation focused on utilizing filter feature selection. Filter feature selection was used to find the top important variables to the prospectively missing variables. Information Gain (IG) was used to measure importance between variables. The important variables were used by KNN imputation that imputes values for the missing variable. The search for nearest neighbors is then conducted using the list of important selected variables. This approach was called Missing Variable Driven Feature Selection KNN imputation (MVDFS-KNN). This approach was tested on varying levels of missingness, starting with one missing variable, and up to five missing variables. Results were compared to the results using the original dataset, mean imputation, traditional KNN imputation, and imputation conducted using CDFS-KNN imputation.

3.1.1 Data Extraction and Preprocessing

The first step of the research framework starts with extracting datasets. The datasets were obtained from UCI Machine learning repository. The UCI machine learning repository includes numerous datasets used by machine learning researchers for the empirical analysis of machine learning algorithms. It includes a wide selection of datasets with varying characteristics. After data installation and reading data files, the data was preprocessed, which includes cleaning, deleting errors, and processing variable names. Then, the data was randomly split into training (80%) and validation (20%) as this data split is a common practice in the literature. The validation set will act as a simulation of a prospective dataset that is missing some variables. Multiple algorithms were evaluated using cross validation to select the best performing one. Algorithms selected in previous literature using the same datasets were considered.

In cross validation, k-folds of data are created where k-1 folds of the dataset are used in model training and the remaining fold is used for testing. This process is iterated k times. Figure 3.2. shows an example of 5-fold cross validation.



Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Figure 3.2: An example of 5-fold cross validation

Cross validation allows for all available data to be used for both training and testing. It is an internal model validation technique that helps detect model overfitting and provides a better indication on model performance on unseen datasets.

3.1.2 Feature Importance Calculation

In the second step, the training set identified in step 1 is used to train and test a prediction model to predict the target. After that, wrapper feature selection is applied to identify a subset of the most important variables. Wrapper feature selection focuses on feature subsets that improve the accuracy of the model in predicting the class/target. It also considers potential interactions between independent variables. In this dissertation, this feature selection is called Class-Driven Feature Selection (CDFS) because it focuses on the accuracy of predicting the class variable. The selected features V_i, $where\ i =$

$1,2, . . , n$, will be saved as a search space that will be used by the KNN algorithm, n is the total number of features selected. Figure 3.3 illustrates how the best feature subset is selected.

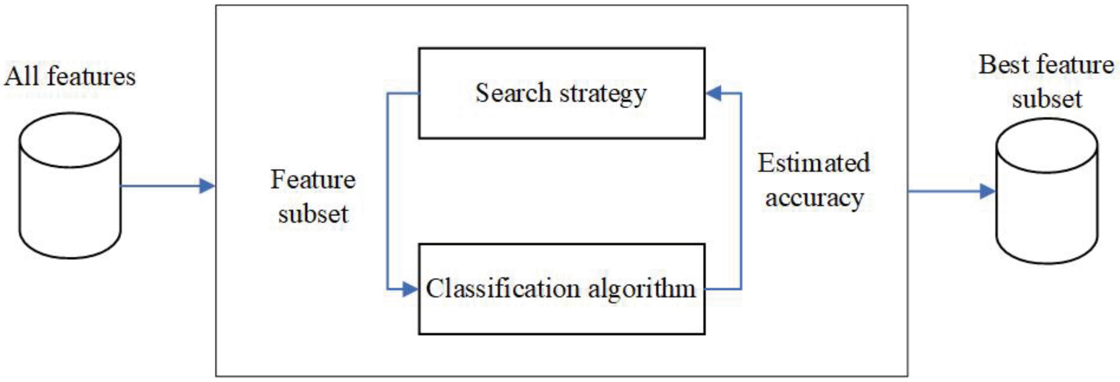


Figure 3.3: Wrapper feature selection to get CDFS features that will be used by KNN imputation

Wrapper feature selection iterates over multiple feature subsets and tests the model accuracy for each feature subset. The feature subset that results in the best accuracy is selected and saved in data frame.

After the important features are identified, a filter feature selection method is applied to identify the top important features for each feature in V_i, $where\ i = 1,2, . . , n$. Filter feature selection methods do not depend on models' performance for variables' subsets. This method uses model independent metrics to calculate association and importance between variables. Information Gain (IG) was used to select the top important variables for each prospectively missing variable. This feature selection method is called Missing Variable-Driven Feature Selection (MVDFS) because it selects the important variables relative to the prospectively missing variable. Figure 3.4 illustrates how filter feature selection works to select important variables.

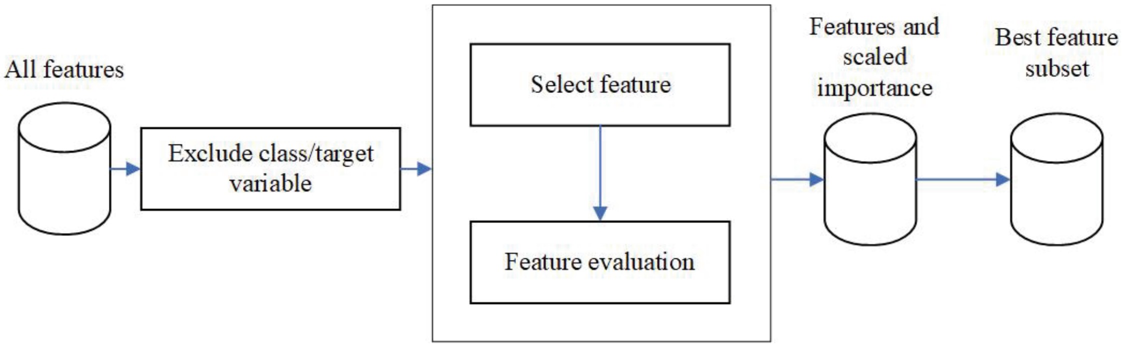


Figure 3.4: Filter feature Selection to get MVDFS features that will be used by KNN imputation.

The IG is a measure that is commonly used in decision trees to determine the variable used to split the input dataset on each node in the tree. Entropy is another measure that determines the impurity in a group of observations. When all the observations in a dataset belong to the same class, the entropy value is 0. This means that the dataset cannot be used for learning. An entropy value of 1 means that the observations are equally distributed across classes. For a dataset with a binary class, the formula for calculating the entropy is as follows:

$$entropy(S) = -p_p\left(log_2\, p_p\right) - p_n(log_2\, p_n)$$

S is the dataset, p_p is the probability of randomly selecting an observation that belongs to the positive class, and p_n is the probability of randomly selecting an observation that belongs to the negative class. The IG formula is explained below.

$$Gain(S, A) = entropy(S) - \sum_{v\, \in Values(A)} \frac{|S_v|}{|S|} entropy(S_v)$$

A represents a variable in the dataset, v represents an element of values in the variable A. $|S_v|$ is the frequency of v in A, $|S|$ is the total number of observations in the dataset. The smaller the attribute's entropy, the higher the IG and the more important it is. The IG was used to measure the importance between all independent variables, then variables were ordered by the IG value in descending order. The variables with IG value greater than or equal to 0.5 were selected as important variables for each variable in V_i, $where\ i = 1,2,..,n$.

3.1.3 Data Missingness Simulation and Imputation

The validation set (20%), identified in step 1, is used in the third step to mimic a prospective dataset. Important variables are censored in the validation set to mimic prospective missingness. Censoring is done in steps, which begins with one missing variable, which is the top important variable, then incrementally increases until five missing important variables. Therefore, there are a total of five cases considered in this research.

For each case, four different methods are used to impute the missing variable: CDFS-KNN, MVDFS-KNN, Mean imputation, and KNN. Additional details about these methods will be provided in section 3.3.

3.1.4 Evaluation and Assessment

In this step, the validation complete dataset is fed to the predictive model to measure the baseline performance of the model on the complete dataset. This includes sensitivity, specificity, Area Under the Receiver Operating Characteristic (ROC) Curve (AUC), and F1 score. All these metrics are extracted from the confusion matrix shown in Table 3.1. The formulas of the metrics are shown in Table 3.2, and the ROC curve is explained in Figure 3.5.

Table 3.1: Confusion matrix

Total Population	**Actual/ True Condition**

		Condition Positive (CP)	**Condition Negative (CN)**
Predicted Condition	**Predicted Condition Positive**	True Positive (TP)	False Positive (FP) Type I error
	Predicted Condition Negative	False Negative (FN) Type II error	True Negative (TN)

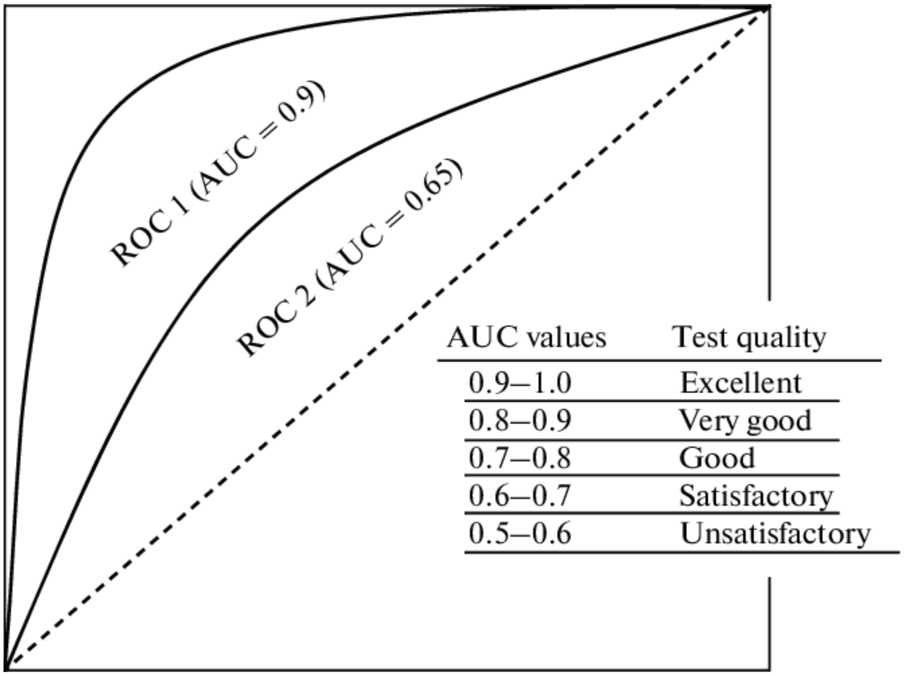
Table 3.2: Classification model evaluation metrics

Metric/ Measure	**Formula**
True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection, Power	$\frac{\sum TP}{\sum CP}$
True Negative Rate (TNR), Specificity (SPC), Selectivity	$\frac{\sum TN}{\sum CN}$
Positive Predictive Value (PPV), Precision	$\frac{\sum TP}{\sum Predicted\ CP}$
F1 Score	$2\frac{PPV.TPR}{PPV+TPR}$

Sensitivity is used to measure the percentage of positive observations predicted and classified correctly by the classifier. Data scientists focus on this metric when the goal of the model is to capture and classify as many positive observations as possible. Sensitivity is also used along with specificity, which measures the percentage of negative observations correctly predicted out of all negative observations.

The F1 score is an overall metric that evaluates the model's performance in terms of the number of correct classifications and how many observations missed. This score reflects the trade-off between sensitivity and precision of a classification model. A higher F1 score denotes better model performance.

The Receiver-Operating characteristics Curve (ROC) is one of the popular curves that are used in measuring the performance of classification models. This curve plots the relationship between False Positive Rate (FPR) and TPR. The area under this curve, AUC, is a very commonly used metric in evaluating classification models. As seen in Figure 3.5, an AUC value of 1 indicates an excellent model, and an AUC value of 0 indicates a poor model. When AUC is 0.5 it means that the model has no class separation.



AUC values	Test quality
0.9–1.0	Excellent
0.8–0.9	Very good
0.7–0.8	Good
0.6–0.7	Satisfactory
0.5–0.6	Unsatisfactory

Figure 3.5: ROC Curve with different AUC values (Trifonova et al., 2013)

The imputed datasets were fed to the classification model one at a time and results were recorded. The results of each model performance metric for each dataset were averaged and plotted, which includes confidence intervals. Each experiment was conducted for multiple iterations to allow for statistical comparisons and obtaining p-values.

3.2 Feature Importance and Feature Selection

Feature importance determines the importance of a variable in a prediction model. It is calculated by measuring the increase in the model's prediction error after permuting a variable. For example, a variable has very low importance when permutation results in no change in the model's prediction error. This implies that the prediction of the model's target is not impacted by this variable. Model agnostic methods rely on feature importance to investigate the amount of contribution variables have on prediction and rank their contributions from the highest to the lowest value. The most important variable permutation results in the greatest change in prediction error (Urbanowicz et al., 2018).

The calculation of feature importance allows for interpretation of how prediction scores are generated, and it provides a global, high-level insight into the models' behavior. It helps in selecting a subset of the top important variables that provides the model with sufficient performance. This approach reduces the input data dimensionality, model complexity, processing time, and computational expenses, when models are deployed to provide predictions. It also reduces the chances of overfitting in the models (Urbanowicz et al., 2018).

Feature selection is a technique in which variables are ranked using various approaches based on their contribution to a model in predicting a target, and then a subset is selected that delivers sufficient accuracy to predict a target. Feature selection methods can be grouped into the following categories: filter methods, wrapper methods, hybrid methods, and embedded methods (Jović et al., 2015). Filter methods are fast and scalable, independent of classifier, and have less risk of overfitting. On the other hand, filter methods do not model dependencies or interactions with classifier. Wrapper methods are classifier dependent, and model feature dependencies. Wrapper methods are slower than filter and embedded methods, and they are more prone to over fitting. Hybrid feature selection is a combination between filter and wrapper methods. It combines the pros and cons of both filter and wrapper methods. Embedded feature selection is focused on selecting the top important features within the model building process. Embedded methods are model dependent and can model feature dependencies. They are faster than wrapper methods, but slower than filter methods (Pudjihartono et al., 2022).

These categories can be applied to either individual or a subset of variables (Bolón-Canedo et al., 2013). Individual evaluation evaluates every feature's relevance to the target and assigns a rank/score to it. Subset evaluation is conducted on a group of features that are selected randomly or based on a search strategy. Initial subsets can be selected randomly; the search strategy involves moving in the direction of improved model performance achieved by the feature subset.

Filter feature selection is done by calculating features' relevance to the target variable through the characteristics of training data. These methods are generally faster and independent from the prediction algorithm. The selected variables can be passed to any prediction algorithm, because their selection criteria do not depend on a classifier performance (Dash and Liu, 1997).

Algorithm 1: Filter feature selection pseudo code

```
Input:
D: A dataset where each row represents an instance, each column represents a feature ($X_1$, $X_2$, …, $X_n$), where n is the total number of features in D
Y: Target variable
Importance between $X_i$ and Y: $f(X_i, Y) = importance(X_i, Y)$
K: number of features selected
```

Output:

Selected features dataset D’

Process:

```
For each $X_i$ in D:
    Calculate $importance\ between\ X_i\ and\ Y, f(X_i, Y)$
Sort values of importance , $f(X_i, Y)$ in a descending order
Select top K features
Save selected features dataset D’
```

In Algorithm 1, the importance is a general term that calculates the association between two variables. For example, the importance can be obtained using Pearson’s correlation, Chi-squared test, Fisher’s exact test, Information gain, and t-test (Pudjihartono et al., 2022). In this research, the Information Gain (IG) was used to calculate importance of features in a dataset as it is one of the common mutual information metrics used in filter feature selection methods. The formula used to calculate IG was provided in section 3.1.2.

Wrapper methods use predictive algorithms to train a model using candidate variables, which follows an iterative process of holding out a variable or a subset of variables to test the model’s performance without those held out (Menze et al., 2009). In each iteration, a model is trained on a subset of variables, which is a reason why applying wrapper methods is computationally intensive. The selection of variables can be done in a forward or backward manner; eliminate one variable at a time, or add one variable at a time, the same logic applies to subset evaluations. Wrapper feature scores are dependent on the prediction algorithm. Scores include classifier accuracy for each subset, the mean decrease in the accuracy or the Gini Index. Therefore, features selected by Random Forest (RF) can be slightly different than features selected by Support Vector Machine (SVM).

Algorithm 2: Wrapper feature selection pseudo code

```
Input:
D: A dataset where each row represents an instance, each column represents a feature ($X_1$, $X_2$, …, $X_n$), where n is the total number of features in D
Y: Target variable
Model: prediction model using features in D to predict Y
Accuracy: accuracy of prediction model
S: multiple subsets of features in D ($S_1$, $S_2$, …, $S_i$), where $i$ is the total number of possible subsets of features

Output:
Selected features dataset D'

Process:
For each S in ($S_1$, $S_2$, …, $S_i$):
        Fit Model to predict Y
        Calculate Accuracy of Model
Sort values of accuracy of each model in a descending order
Select feature subset with highest accuracy value
Save selected features D'
```

In embedded feature selection, finding the optimal feature subset is done as a part of training the prediction model. These methods select features by optimizing two objective functions: the number of features selected and classification accuracy. Embedded methods and wrapper methods are both dependent on a prediction algorithm to predict the target variable. However, the difference is that embedded methods select features without reclassifying different feature subsets as in wrapper methods. This reduces the computation time compared to wrapper methods (Chen et al., 2020).

3.3 Missing Data Imputation

Datasets that have missing values can impact the process of training and testing predictive models. Missing values introduce bias to models, increase the chances of errors, and could significantly impact model outcomes and conclusions (Kang, 2013). The presence of missing data impacts the deployment of these models, which was discussed in Chapter 1. There are multiple methods for handling data missingness as a part of preprocessing the data before model development.

Missing data is handled through multiple methods; deletion is one of the common methods researchers apply to datasets. Complete case analysis, also known as case-wise deletion, where records with missing values are deleted is an example of the deletion method. This method is commonly used in treatment studies where no assumptions are made about the missing data. The disadvantage of this method is that it reduces the sample size of the dataset, which could provide information to the analysis conducted if included.

Other approaches that have been used to handle missing data include imputation-based and non-imputation-based methods. This categorization of missing data handling methods was proposed in the literature (Nijman et al., 2022). Imputation-based methods include single or multiple imputation methods that would fill in a missing value with a plausible estimate of that data point. This includes mean and mode imputation, KNN imputation, and other multiple imputation methods. Non-imputation-based methods learn from missingness patterns and do not fill in values to replace missing data. This includes machine learning models and likelihood-based methods such as expectation-maximization.

KNN imputation maintains the variability of the datasets compared to mean and mode imputation. KNN imputation was reported to be efficient, and flexible in both discrete and continuous data (Kang, 2013). KNN imputation does not require the construction of a prediction model for each input variable that has missing values, unlike model-based methods. It learns from similar complete observations in the training set. KNN imputation has shown outperforming results when compared to other methods such as decision trees, mean, and mode imputation. The KNN imputation algorithm performs well even when a large amount of missing data is present (Batista and Monard, 2003, De Silva and Perera, 2016). Although this method has been highly researched in the literature, the use of KNN imputation in the big data setting is still an underexplored area (Emmanuel et al., 2021).

3.3.1 KNN Imputation

KNN is a common supervised classification algorithm that calculates distances between a data point and other data points. After that, it selects the nearest K observations to vote for the most frequent class to classify that data point of interest (Peterson, 2009). Figure 3.6 presents a KNN classification example. The yellow point represents a new observation that can be classified either as Class 1 or Class 2. When the three nearest neighbors are selected, the majority vote is for Class 2. However, when the number of neighbors is increased to seven, the majority vote is for Class 1. The commonly used distance measure is the normalized Euclidean distance (Cohen, 2016). Euclidean distance was used in KNN imputation methods in this dissertation.

For classification problems, the majority vote of nearest neighbors is used to classify the data point. For regression problems or continuous targets, the mean or median of the nearest neighbors is the most common approach that is used to calculate the target value for the data point.

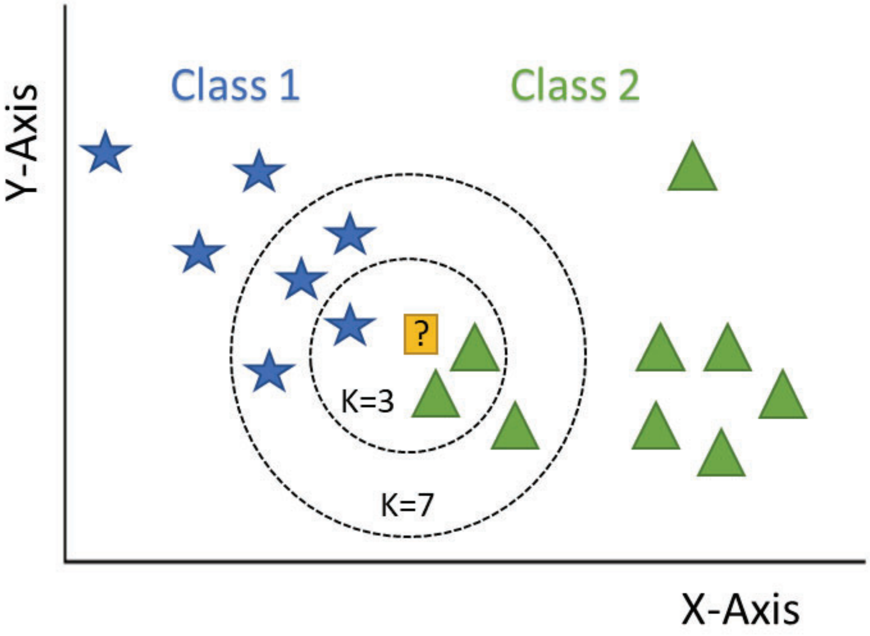


Figure 3.6: Classification example using KNN algorithm

KNN imputation (Andridge and Little, 2010) uses the same concept to impute missing values from the nearest observations. In this imputation method, missing values are replaced by values extracted from observations that are similar to the data point with missing values concerning observed characteristics (Beretta and Santaniello, 2016). The imputed value is either an exact measured value (one nearest neighbor) or an average of measured values of k neighbors. When datasets are being preprocessed before model development, KNN imputation is applied to fill in the gaps where each observation with missing value(s) is fed into the KNN algorithm to search for similar observations. All the independent variables (model inputs) are used to find the nearest neighbors before the average or majority vote is calculated. The target or class value is not considered when searching for nearest neighbors as it could potentially introduce bias in the predictive model. The steps of KNN imputation are provided in Algorithm 3.

KNN imputation replaces missing values with actual measured values and not constructed values. It utilizes auxiliary information provided by other variables, which preserves the dataset’s original structure (Beretta and Santaniello, 2016).

Algorithm 3: KNN imputation pseudo code

```
Input:
Upload required dataset D = {$x_1, x_2, \ldots, x_n$}; where x represents features, n is the number of features.
Set number of neighbors K

Output:
Complete dataset D’

Process:
For each missing value in D:
    Find k-nearest neighbors of the observation with missing value
    Calculate average (or median or mode) of the feature values for k-nearest neighbors
    Replace missing value with calculated average (or median or mode)
Save complete dataset D’
```

3.3.2 CDFS-KNN Imputation

The CDFS-KNN imputation method combines wrapper feature selection with traditional KNN imputation. As discussed earlier, missing data on deployment impacts the ability to implement models prospectively. This technique allows the deployment of models on prospective data even when variables are missing or not available at the time of deployment. Mean imputation and KNN imputation can be applied to handle prospectively missing variables. However, some of the previous studies in the literature showed that mean imputation and KNN could cause inaccurate risk predictions and therefore model accuracy drop. Algorithm 4 illustrates the steps of CDFS-KNN imputation.

Algorithm 4: CDFS-KNN imputation pseudo code

```
Input:
$D_t$: A training dataset where each row represents an instance, each column represents a
feature ($X_1$, $X_2$, …, $X_n$), where n is the total number of features in $D_t$
$D_p$: A prospective dataset where each row represents an instance, each column represents a
feature ($V_1$, $V_2$, …, $V_n$), where n is the total number of features in $D_p$
m: Number of missing features in $D_p$
Y: Target variable
Model: prediction model using features in $D_t$ to predict Y
S: Multiple subsets of features in $D_t$ ($S_1$, $S_2$, …, $S_i$), where i is the total number of possible
subsets
K: Number of neighbors

Output:
I: List of important features
$D_p$': Complete prospective dataset

Process:
For each S in ($S_1$, $S_2$, …, $S_i$):
        Fit Model to predict Y
        Calculate Accuracy (Model)
Sort values of accuracy of each subset in a descending order
Select subset of features with highest accuracy value
Save selected features I
For each missing variable in $D_p$:
        Select I in $D_t$ and $D_p$
        Use I in $D_t$ to Find k-nearest neighbors to the observation with missing values in $D_p$
        Calculate average (or mode) of the feature values for k-nearest neighbors
        Replace missing value in $D_p$ with calculated average (or mode) from $D_t$
```

```
Save complete dataset D_p'
```

Selecting the top important variables using wrapper feature selection is completed first. These variables identify the n-dimensional search space for the KNN imputation algorithm, where n is the number of selected variables. Then, when KNN is applied to impute prospective missing variables, the algorithm reaches to the n-dimensional search space to find nearest neighbors. Different number of neighbors will be used to study the impact of the number of neighbors on the performance CDFS-KNN. The higher the number of neighbors used, the higher the computational expenses are to process the neighbor search and the imputation.

3.3.3 MVDFS-KNN Imputation

This method starts with selecting the top important variables for each variable in $V_i\ , where\ i \in [1, m]$. This is done using filter feature selection that uses the IG as a feature importance measure. For each missing variable in the validation/prospective set, the search space for the KNN imputation algorithm is identified through its relative important variables. For each variable to be missing prospectively, a list of important variables is selected using filter feature selection. Searching for nearest neighbors to do the imputation is done utilizing only the important features, not all features in the training dataset. This methodology utilizes the relationship of the missing variable with the other input variables to impute missing values. Algorithm 5 illustrates the steps of MVDFS-KNN imputation, which combines filter feature selection using IG and the KNN imputation.

Algorithm 5: MVDFS-KNN imputation pseudo code

```
Input:
$D_t$: A training dataset where each row represents an instance, each column represents a feature ($X_1$, $X_2$, …, $X_n$), where n is the total number of features in $D_t$
$D_p$: A prospective dataset where each row represents an instance, each column represents a feature ($V_1$, $V_2$, …, $V_n$), where n is the total number of features in $D_p$
m: Number of missing features in $D_p$
Y: Target variable
Information Gain between $X_i$ and Y: $f(X_i, Y) = IG(X_i, Y)$
K: Number of neighbors
Output:
I: List of important features for each $V_i$ in ($V_1$, $V_2$, …, $V_m$)
$D_p$': Complete prospective dataset
Process:
For each missing $V_i \in$ ($V_1$, $V_2$, …, $V_m$) in $D_p$:
    Calculate $IG\ between\ X_i\ and\ and\ other\ variables\ excluding\ Y, f(X, Y) = IG(X_i, Y)$
    Sort values of $IG$ in a descending order
    Select top features with $IG \geq 0.5$
    Save selected important features $I_{Vi}$
For each missing variable $V_i \in$ ($V_1$, $V_2$, …, $V_m$) in $D_p$:
    Select $I_{Vi}$ in $D_t$ and $D_p$
    Use I to Find k-nearest neighbors to the observation with missing values in $D_p$
    Calculate average (or mode) of the feature values for k-nearest neighbors
    Replace missing value in $D_p$ with calculated average (or mode) from $D_t$
Save complete dataset $D_p$'
```

The higher the number of neighbors used in MVDFS-KNN, the higher the computational expenses are to process the neighbor search and the imputation. A different number of neighbors will be used to study the impact of the number of neighbors on the performance MVDFS-KNN.

3.4 Chapter Summary

This chapter provides a detailed explanation of the research framework and steps followed to achieve results. It starts with data extraction and preprocessing, followed by selecting important features using two approaches: wrapper feature selection, and filter feature selection using IG. The wrapper feature selection approach focuses on the model's accuracy in predicting the target/class, it is named CDFS. The filter feature selection is model independent. It was utilized to select the important independent features with respect to features that are expected to be missing prospectively. This approach is named MVDFS.

The methodology starts with extracting and preprocessing the dataset. This includes reading the documentation of each dataset to understand variables' definitions and how to handle errors. Then data preprocessing, which includes cleaning, deleting errors, and processing variable names, was conducted. Then, the data was randomly split into training (80%) and validation (20%) sets. The validation set will act as a simulation of a prospective dataset that is missing some variables.

Multiple algorithms were evaluated using cross validation to select the best one. Algorithms selected in previous literature using the same datasets were considered. After selecting the prediction algorithm, wrapper feature selection was used to select the top important variables with respect to the model's accuracy in predicting the class. This technique is called Class Driven Feature Selection (CDFS), because it selects the important feature for predicting class. The important variables were saved in a data frame V_i, $where\ i = 1,2,..,n$.

After the important features are identified, a filter feature selection method is applied to identify the top important features for each feature in V_i, $where\ i = 1,2,..,n$.

Information Gain (IG) was used to select the top important variables for each prospectively missing variable. This feature selection method is called Missing Variable-Driven Feature Selection (MVDFS) because it selects the important variables relative to the prospectively missing variable. CDFS-KNN imputation combined CDFS and KNN imputation, where the search for nearest neighbors was done using the variables selected only by the CDFS approach. MVDFS-KNN imputation combined MVDFS and KNN imputation, where the search for nearest neighbors for each missing variables was done using the variables selected by the MVDFS approach.

Important variables were censored in the validation set to mimic prospective missingness. Censoring was done in steps; it began with one missing variable, which is the top important variable, then incrementally increases until five important variables are missing. Therefore, there are a total of five cases considered. Five datasets were entered to the model to record model performance metrics. These datasets are the original validation set, missing variables imputed by MI, imputed by KNN imputation, imputed by CDFS-KNN, and by MVDFS-KNN. Results including F1 score, AUC, Sensitivity, and Specificity were documented for each imputation approach and each dataset. This process was repeated for multiple iterations to allow for statistical hypothesis testing.

Chapter 4: Experimental Results and Analysis

This chapter discusses the results of applying the CDFS-KNN and MVDFS-KNN imputation on five datasets. Section 4.1 discusses the datasets used in this dissertation and the resource they were extracted from. It provides a detailed description of the datasets and reasons for including them in this research. Section 4.2 includes the experimental results for the five datasets. This includes model performance on original datasets and imputed datasets using MI, KNN imputation, CDFS-KNN imputation, and MVDFS-KNN imputation. Section 4.3 provides a summary of the experimental results.

4.1 Data Description

Multiple public datasets were collected from the UCI Machine Learning Repository (Asuncion and Newman, 2007). The datasets are static and were collected under a static timeframe. The datasets were selected to test the applicability and performance of the proposed methods on datasets with different number of instances, dimensionality, variable types, and balance levels between minority and majority classes. Table 4.1 summarizes the five datasets selected, the number of features, the number of instances, and the class balance represented by percentage of the positive class. All the datasets listed have a binary response variable.

Table 4.1: Description of datasets

Dataset	**Source**	**Number of Features**	**Number of Instances**	**Class Balance (%Positive)**
Thoracic Surgery (TS)	(Zieba et al., 2014)	17	470	14%

Breast Cancer Wisconsin (BCW)	(Wolberg et al., 1995)	32	569	37%
Cervical Cancer (CC)	(Fernandez et al., 2017)	36	858	12%
Adult	(Kohavi, 1996)	15	32,561	24%
Diabetes 130 US Hospitals	(Strack et al., 2014)	51	101,766	11%

The Thoracic Surgery dataset includes 17 variables collected about lung cancer patients who underwent major lung resections for primary lung cancer between the years 2007 and 2011. The variables include demographic information, symptoms before surgery, chronic diseases, diagnoses, and test-related variables. Table 4.2 summarizes the variable types and their description. The dataset is used to predict patients' life expectancy. The target variable has two classes: class 1 represents death within first year after surgery; class 2 represents patients who survived. The dataset has a small number of instances, and it is imbalanced with a class balance percentage of 14%.

Table 4.2: Thoracic Surgery dataset attributes description (Zieba et al., 2014)

ID	Attribute Name	Attribute Description	Attribute Type	Attribute Range or Values
1	DGN	Diagnosis—specific combination of ICD-10 codes for primary and secondary as well multiple tumors if any (DGN3, DGN2, DGN4, DGN6, DGN5, DGN8, DGN1)	Nominal	DGN1: 1, DGN2: 2, DGN3: 3, DGN4: 4, DGN5: 5, DGN6: 6, DGN8: 8
2	PRE4	Forced Vital Capacity (FVC)	Numeric	1.44-6.30
3	PRE5	Volume that has been exhaled at the end of the first second of forced expiration (FEV1)	Numerical	0.96-86.3
4	PRE6	Performance status—Zubrod scale (PRZ2, PRZ1, PRZ0)	Nominal	PRZ0: 0, PRZ1: 1, PRZ2: 2
5	PRE7	Presence of pain before surgery	Nominal	F: 0, T: 1
6	PRE8	Presence of hemoptysis (coughing up blood)	Nominal	F: 0, T: 1

7	PRE9	Presence of dyspnea (difficulty breathing)	Nominal	F: 0, T: 1
8	PRE10	Presence of cough	Nominal	F: 0, T: 1
9	PRE11	Presence of weakness	Nominal	F: 0, T: 1
10	PRE14	T in clinical TNM—size of the original tumor, from OC11 (smallest) to OC14 (largest) (OC11, OC14, OC12, OC13)	Nominal	OC11: 1 OC12: 2 OC13: 3 OC14: 4
11	PRE17	Type 2 DM—diabetes mellitus	Nominal	F: 0, T: 1
12	PRE19	MI up to 6 months	Nominal	F: 0, T: 1
13	PRE25	PAD—peripheral arterial diseases	Nominal	F: 0, T: 1
14	PRE30	Smoking	Nominal	F: 0, T: 1
15	PRE32	Asthma	Nominal	F: 0, T: 1
16	AGE	Age at surgery	Numerical	21-87
17	Risk1Yr	Risk for death in one year after surgery	Nominal	F: 0, T: 1

The Breast Cancer Wisconsin dataset is a diagnostic dataset used for classifying breast masses into two classes: malignant and benign. It includes 32 features computed from a digital image of a Fine Needle Aspirate (FNA) of a mass in the breast. The 32 features measure the mass dimensions and other characteristics. Table 4.3 summarizes the variable types and their description. This dataset was selected because it has a slightly larger number of instances compared to the thoracic surgery dataset and has higher dimensionality. Also, the data is not highly imbalanced where approximately 37% of the dataset belongs to malignant class.

Table 4.3: Breast Cancer Wisconsin dataset attributes description (Wolberg et al., 1995)

ID	Attribute Name	Attribute Description	Attribute Type	Attribute Range or Values
1	ID	Unique identification number for each case	Nominal	N/A
2	Diagnosis	Diagnosis of cancer (M for Malignant, B for Benign)	Nominal	M (Malignant), B (Benign)
3	Radius_mean	Mean of distances from center to points on the perimeter	Numerical	6.981-28.11
4	Texture_mean	Mean value for standard deviation of gray-scale values	Numerical	9.71-39.28
5	Perimeter_mean	Mean size of the core tumor mass's boundary	Numerical	43.79-188.5

6	Area_mean	Mean area of the core tumor mass	Numerical	143.5-2501.0
7	Smoothness_mean	Mean of local variation in radius lengths	Numerical	0.053-0.163
8	Compactness_mean	Mean of perimeter^2 / area - 1	Numerical	0.019-0.345
9	Concavity_mean	Mean of severity of concave portions of the contour	Numerical	0.0-0.427
10	Concave points_mean	Mean for the number of concave portions of the contour	Numerical	0.0-0.201
11	Symmetry_mean	Mean of symmetry	Numerical	0.106-0.304
12	Fractal_dimension_mean	Mean of "coastline approximation" - 1	Numerical	0.05-0.097
13	Radius_se	Standard error of the mean of distances from center to points on the perimeter	Numerical	0.112-2.873
14	Texture_se	Standard error of the mean value for the standard deviation of gray-scale values	Numerical	0.36-4.885
15	Perimeter_se	Standard error of the mean size of the core tumor mass's boundary	Numerical	0.757-21.98
16	Area_se	Standard error of the mean area of the core tumor mass	Numerical	6.802-542.2
17	Smoothness_se	Standard error of the mean of local variation in radius lengths	Numerical	0.002-0.031
18	Compactness_se	Standard error of the mean of perimeter^2 / area - 1	Numerical	0.002-0.135
19	Concavity_se	Standard error of the mean of the severity of concave portions of the contour	Numerical	0.0-0.396
20	Concave points_se	Standard error of the mean for the number of concave portions of the contour	Numerical	0.0-0.053
21	Symmetry_se	Standard error of the mean of symmetry	Numerical	0.008-0.079
22	Fractal_dimension_se	Standard error of the mean of "coastline approximation" - 1	Numerical	0.001-0.03
23	Radius_worst	Worst (largest) value of distances from center to points on the perimeter	Numerical	7.93-36.04
24	Texture_worst	Worst (largest) value for standard deviation of gray-scale values	Numerical	12.02-49.54
25	Perimeter_worst	Worst (largest) size of the core tumor mass's boundary	Numerical	50.41-251.2

26	Area_worst	Worst (largest) area of the core tumor mass	Numerical	185.2-4254.0
27	Smoothness_worst	Worst (largest) variation in radius lengths	Numerical	0.071-0.223
28	Compactness_worst	Worst (largest) value of perimeter^2 / area - 1	Numerical	0.027-1.058
29	Concavity_worst	Worst (largest) severity of concave portions of the contour	Numerical	0.0-1.252
30	Concave points_worst	Worst (largest) number of concave portions of the contour	Numerical	0.0-0.291
31	Symmetry_worst	Worst (largest) symmetry value	Numerical	0.156-0.664
32	Fractal_dimension_worst	Worst (largest) "coastline approximation" - 1	Numerical	0.055-0.208

The remaining three datasets were selected because they have higher dimensionality, larger number of instances, and a varying degree of class imbalance. The Cervical Cancer dataset was selected because it has similar dimensionality compared to the Breast Cancer Wisconsin data and the number of instances is slightly higher than that of the BCW dataset. The dataset is imbalanced where 12% of the instances were diagnosed with cervical cancer. Table 4.4 summarizes the variable types and their description.

Table 4.4: Cervical Cancer dataset attributes description (Fernandez et al., 2017)

ID	Attribute Name	Attribute Description	Attribute Type	Attribute Range or Values
1	Age	Age of the patient (in years)	Numerical	13-84 years
2	Number of sexual partners	Number of sexual partners	Numerical	0-28
3	First sexual intercourse	Age at first sexual intercourse (in years)	Numerical	10-32 years
4	Num of pregnancies	Number of pregnancies	Numerical	0-11
5	Smokes	Smoking status	Nominal	0: No, 1: Yes
6	Smokes (years)	Number of years of smoking (if applicable)	Numerical	0-37 years
7	Smokes (packs/year)	Number of packs of cigarettes per year	Numerical	0-37 packs/year
8	Hormonal contraceptives	Use of hormonal contraceptives	Nominal	0: No, 1: Yes

9	Hormonal contraceptives (years)	Number of years of hormonal contraceptive use (if applicable)	Numerical	0-30 years
10	IUD	Use of intrauterine device	Nominal	0: No, 1: Yes
11	IUD (years)	Number of years of intrauterine device use (if applicable)	Numerical	0-19 years
12	STDs	History of sexually transmitted diseases	Nominal	0: No, 1: Yes
13	STDs (number)	Number of sexually transmitted diseases (if applicable)	Numerical	0-4
14	STDs:condylomatosis	Presence of condylomatosis	Nominal	0: No, 1: Yes
15	STDs:cervical condylomatosis	Presence of cervical condylomatosis	Nominal	0: No, 1: Yes
16	STDs:vaginal condylomatosis	Presence of vaginal condylomatosis	Nominal	0: No, 1: Yes
17	STDs:vulvo-perineal condylomatosis	Presence of vulvo-perineal condylomatosis	Nominal	0: No, 1: Yes
18	STDs:syphilis	Presence of syphilis	Nominal	0: No, 1: Yes
19	STDs:pelvic inflammatory disease	Presence of pelvic inflammatory disease	Nominal	0: No, 1: Yes
20	STDs:genital herpes	Presence of genital herpes	Nominal	0: No, 1: Yes
21	STDs:molluscum contagiosum	Presence of molluscum contagiosum	Nominal	0: No, 1: Yes
22	STDs:AIDS	Presence of AIDS	Nominal	0: No, 1: Yes
23	STDs:HIV	Presence of HIV	Nominal	0: No, 1: Yes
24	STDs:hepatitis B	Presence of Hepatitis B	Nominal	0: No, 1: Yes
25	STDs:HPV	Presence of HPV	Nominal	0: No, 1: Yes
26	STDs: number of diagnoses	Number of STD diagnoses	Numerical	0-3
27	STDs: time since first diagnosis	Time since the first STD diagnosis (in years)	Numerical	1-22
28	STDs: time since last diagnosis	Time since the last STD diagnosis (in years)	Numerical	1-22
29	Dx:cancer	Presence of cervical cancer	Nominal	0: No, 1: Yes
30	Dx:CIN	Presence of cervical intraepithelial neoplasia (CIN)	Nominal	0: No, 1: Yes
31	Dx:HPV	Presence of HPV infection	Nominal	0: No, 1: Yes
32	Dx	Presence of any cervical disease	Nominal	0: No, 1: Yes
33	Hinselmann	Presence of Hinselmann test	Nominal	0: No, 1: Yes
34	Schiller	Presence of Schiller test	Nominal	0: No, 1: Yes
35	Citology	Presence of cytology test	Nominal	0: No, 1: Yes
36	Biopsy	Presence of biopsy	Nominal	0: No, 1: Yes

The Adult dataset was selected because it has a larger number of instances, but it has low dimensionality. This dataset was collected from the 1994 Census database to predict whether individuals' income is less than or greater than $50K/year. The attributes include demographic, education-related, and income-related information. The dataset is relatively balanced where 24% of the instances have income greater than $50K/ year. Table 4.5 summarizes the variable types and their description.

Table 4.5: Adult dataset attributes description (Kohavi, 1996)

ID	Attribute Name	Attribute Description	Attribute Type	Attribute Range or Values
1	Age	Age of the individual	Numerical	17-90 years
2	Workclass	Type of employment or workclass	Nominal	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
3	Fnlwgt	Final weight; sampling weight indicating population	Numerical	12285-1484705
4	Education	Highest level of education achieved	Nominal	Bachelors, Masters, Doctorate, Prof-school, Assoc-acdm, Assoc-voc, Some-college, HS-grad, 12th, 11th, 10th, 9th, 7th-8th, 5th-6th, 1st-4th, Preschool
5	Education-Num	Numeric representation of education level	Numerical	1-16
6	Marital-Status	Marital status of the individual	Nominal	Never-married, Divorced, Separated, Widowed, Married-civ-spouse, Married-spouse-absent, Married-AF-spouse
7	Occupation	Type of occupation	Nominal	Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspection, Adm-clerical, Farming-fishing, Transport-moving, Private-house-serv, Protective-serv, Armed-Forces
8	Relationship	Relationship status of the individual	Nominal	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
9	Race	Race of the individual	Nominal	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
10	Sex	Gender of the individual	Nominal	Female, Male
11	Capital-Gain	Capital gains reported by the individual	Numerical	0-99999
12	Capital-Loss	Capital losses reported by the individual	Numerical	0-4356

13	Hours-Per-Week	Hours worked per week	Numerical	1-99
14	Native-Country	Native country of the individual	Nominal	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-Scotland, Trinidad and Tobago, Greece, Nicaragua, Vietnam, Hong, Ireland, Hungary, Holland-Netherlands
15	Income	Income level	Nominal	<=50K, >50K

The last and largest dataset selected is the Diabetes 130 US hospitals dataset. This dataset has a large number of instances (n=101,766) and has high dimensionality with 51 attributes. The dataset is used to predict patients' readmission in less than or greater than 30 days from discharge for diabetic inpatients. The dataset includes demographic, medical, laboratory results, and hospital stay-related information. It also includes variables that describe medications administered and medical history in the year before hospitalization. Table 4.6 summarizes the variable types and their description. This dataset is imbalanced as readmissions in 30 days or less represent 11% of the instances.

Table 4.6: Diabetes dataset attributes description (Strack et al., 2014)

ID	Attribute Name	Attribute Description	Attribute Type	Attribute Range or Values
1	Encounter ID	Unique identifier of an encounter	Numerical	N/A
2	Patient number	Unique identifier of a patient	Numerical	N/A
3	Race	Patient race	Nominal	Caucasian, Asian, African American, Hispanic, and other
4	Gender	Patient gender	Nominal	Male, Female, and Unknown/Invalid
5	Age	Age on admission	Nominal	10-year intervals: [0, 10), [10, 20), ..., [90, 100)
6	Weight	Weight of patient	Nominal	25-lb intervals: [0,25), [25,50), …, [175, 200), Greater than 200
7	Admission type	Integer identifier corresponding to 8 distinct values representing type of admission	Nominal	Examples: emergency, urgent, elective, newborn, Trauma center, not mapped

8	Discharge disposition	Integer identifier corresponding to 29 distinct values representing discharge disposition	Nominal	Examples: discharged to home, expired, left AMA, and not available
9	Admission source	Integer identifier corresponding to 26 distinct values representing source of admission	Nominal	Examples: physician referral, emergency room, and transfer from a hospital
10	Time in hospital	Number of days between admission and discharge	Numerical	1-14
11	Payer code	Integer identifier corresponding to 23 distinct values representing insurance type	Nominal	Examples: Blue Cross\Blue Shield, Medicare, and self-pay
12	Medical specialty	Integer identifier corresponding to 84 distinct values representing the specialty of the admitting physician	Nominal	Examples: cardiology, internal medicine, family\general practice, and surgeon
13	Number of lab procedures	Number of lab tests performed during the encounter	Numerical	1-132
14	Number of procedures	Number of procedures (other than lab tests) performed during the encounter	Numerical	0-6
15	Number of medications	Number of distinct generic medication names administered during the encounter	Numerical	1-81
16	Number of outpatient visits	Number of outpatient visits of the patient in the year preceding the encounter	Numerical	0-42
17	Number of emergency visits	Number of emergency visits of the patient in the year preceding the encounter	Numerical	0-76
18	Number of inpatient visits	Number of inpatient visits of the patient in the year preceding the encounter	Numerical	0-21
19	Diagnosis 1	The primary diagnosis coded as first three digits of ICD9	Nominal	717 distinct values
20	Diagnosis 2	Secondary diagnosis coded as first three digits of ICD9	Nominal	749 distinct values
21	Diagnosis 3	Additional secondary diagnosis coded as first three digits of ICD9	Nominal	790 distinct values
22	Number of diagnoses	Number of diagnoses entered into the system	Numerical	1-16
23	max_glu_seru m	Indicates the range of the Glucose serum test result or if the test was not taken	Nominal	Values: ">200", ">300", "normal", and "none" if not measured

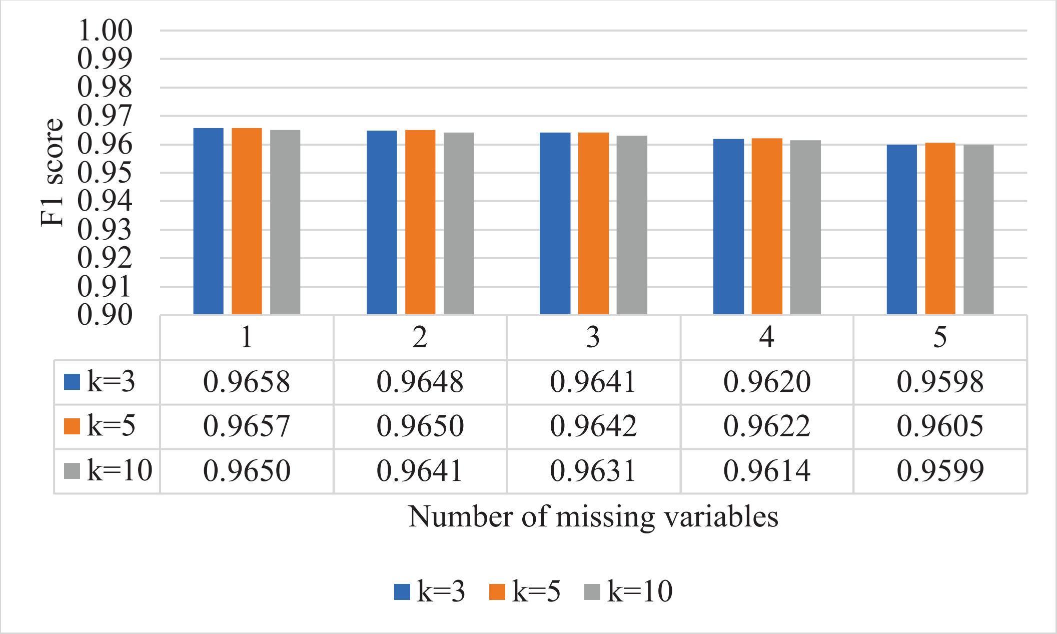
24	A1C test result	Indicates the range of the A1C test result or if the test was not taken	Nominal	Values: ">8": Result greater than 8%, ">7": Result greater than 7% but less than 8%, "normal" Result less than 7%, and "none" if not measured
25	Change of medications	Indicates if there was a change in diabetic medications (either dosage or generic name)	Nominal	Values: "change" and "no change"
26	Diabetes medications	Indicates if there was any diabetic medication prescribed	Nominal	Values: "yes" and "no"
27	24 features for medications	For the generic names: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, the feature indicates whether the drug was prescribed or there was a change in the dosage	Nominal	Values: "up": dosage was increased during the encounter, "down": dosage was decreased, "steady": dosage did not change, and "no": drug was not prescribed
28	Readmitted	Days to inpatient readmission from discharge	Nominal	Values: "<30": readmitted in less than 30 days, ">30" : readmitted in more than 30 days, and "No": no record of readmission

Synthetic Minority Over-sampling Technique (SMOTE) was applied on imbalanced datasets. This technique generates plausible examples from the minority class that are relatively close to the existing examples in the dataset. SMOTE has shown successful results compared to other common data balancing methods based on random sampling with replacement (Abdoh et al., 2018).

4.2 Experimental Results

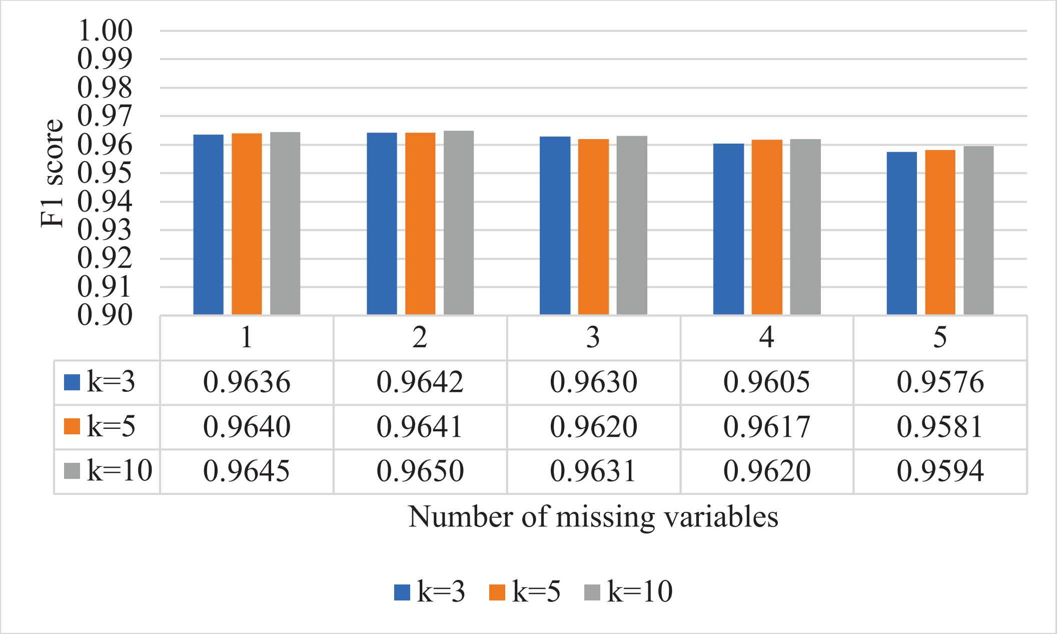
The five datasets were used to test the two proposed approaches: CDFS-KNN and the MVDFS-KNN. The results were compared with common data imputation methods, which included Mean Imputation (MI) and K-Nearest Neighbor (KNN). Twenty percent of each dataset was selected randomly as a validation set. The validation set was used to simulate a prospective dataset with different number of missing variables.

The number of neighbors in the KNN imputation methods was selected as three. This number was identified through experiments with 3, 5, and 10 neighbors using the Breast Cancer Wisconsin dataset. Fifty iterations were conducted to record the model performance on original and imputed datasets among the different number of neighbors. Figures 4.1 and 4.2 show the F1 scores for CDFS-KNN and MVDFS-KNN, respectively, at 3, 5, and 10 neighbors across the number of missing variables.



	1	2	3	4	5
k=3	0.9658	0.9648	0.9641	0.9620	0.9598
k=5	0.9657	0.9650	0.9642	0.9622	0.9605
k=10	0.9650	0.9641	0.9631	0.9614	0.9599

Figure 4.1: F1 scores for CDFS-KNN vs. number of missing variables at 3, 5, and 10 neighbors.



	1	2	3	4	5
k=3	0.9636	0.9642	0.9630	0.9605	0.9576
k=5	0.9640	0.9641	0.9620	0.9617	0.9581
k=10	0.9645	0.9650	0.9631	0.9620	0.9594

Figure 4.2: F1 scores for MVDFS-KNN vs. number of missing variables at 3, 5, and 10 neighbors.

It was determined that the results did not vary significantly across the different number of neighbors for both CDFS-KNN and MVDFS-KNN. Therefore, the number of neighbors was set at 3 for both approaches to reduce computational time. The results of CDFS-KNN and MVDFS-KNN compared to KNN and MI for each of the five datasets are summarized in the following sections.

4.2.1 Thoracis Surgery Data Experiments

A RF model was fitted to predict mortality among thoracic surgery patients. Five-fold cross validation was applied with 50 iterations. The RF algorithm was selected because it resulted in high classification accuracy when applied on the thoracic surgery dataset in recent literature (Ravichandran et al., 2021).

The top selected variables using wrapper feature selection, which are ordered with descending importance were the following: DGN_DGN3, PRE14_OC11, PRE10_T, PRE5, PRE30_T, PRE11_T, PRE6_PRZ0, PRE17_T, DGN_DGN5, and PRE14_OC14. The F1 score, AUC, sensitivity, and specificity of the model were recorded on the original validation dataset (20% of dataset), and on the imputed datasets with one to five missing variables. Table 4.7 summarizes the model performance measures for each imputation method with the number of missing variables. The experiments, repeated for 50 iterations, were used to construct confidence intervals. Figure 4.3 summarizes the average F1 scores for original and imputed datasets vs. number of missing variables.

Table 4.7: Average performance measures of RF for TS original and imputed datasets

Measure	# of Missing Variables	Imputation Methods				Original Dataset
		CDFS-KNN	MVDFS-KNN	KNN	MI	
AUC	1	0.9201 ± 0.0055	0.9170 ± 0.0054	0.8415 ± 0.0121	0.8244 ± 0.0084	0.9242 ± 0.0024
	2	0.9160 ± 0.0057	0.8988 ± 0.0066	0.8416 ± 0.0119	0.7536 ± 0.0074	
	3	0.9160 ± 0.0054	0.8923 ± 0.0073	0.8183 ± 0.0144	0.7257 ± 0.0092	
	4	0.8977 ± 0.0065	0.8666 ± 0.0070	0.8072 ± 0.0143	0.7180 ± 0.0105	
	5	0.8897 ± 0.0063	0.8612 ± 0.0068	0.7797 ± 0.0152	0.6890 ± 0.0163	
Sensitivity	1	0.8379 ± 0.0104	0.8354 ± 0.0112	0.8810 ± 0.0137	0.9274 ± 0.0092	0.8331 ± 0.0045
	2	0.8393 ± 0.0096	0.8420 ± 0.0112	0.8811 ± 0.0141	0.9757 ± 0.0058	
	3	0.8416 ± 0.0090	0.8405 ± 0.0106	0.8920 ± 0.0144	0.9883 ± 0.0038	
	4	0.8266 ± 0.0094	0.8228 ± 0.0107	0.8844 ± 0.0138	0.9791 ± 0.0055	
	5	0.8286 ± 0.0097	0.8255 ± 0.0110	0.8962 ± 0.0144	0.9880 ± 0.0037	
Specificity	1	0.9725 ± 0.0052	0.9691 ± 0.0063	0.7969 ± 0.0264	0.7075 ± 0.0193	0.9803 ± 0.002
	2	0.9660 ± 0.0058	0.9390 ± 0.0089	0.7968 ± 0.0262	0.3554 ± 0.0199	
	3	0.9649 ± 0.0058	0.9301 ± 0.0097	0.7247 ± 0.0364	0.1806 ± 0.0180	
	4	0.9463 ± 0.0082	0.9005 ± 0.0106	0.7097 ± 0.0365	0.2097 ± 0.0214	
	5	0.9335 ± 0.0081	0.8897 ± 0.0108	0.6193 ± 0.0433	0.0997 ± 0.0179	
F1 Score	1	0.8955 ± 0.0062	0.8918 ± 0.0066	0.8295 ± 0.0116	0.8133 ± 0.0099	0.8972 ± 0.0029
	2	0.8927 ± 0.0060	0.8787 ± 0.0071	0.8294 ± 0.0113	0.7091 ± 0.0101	
	3	0.8934 ± 0.0057	0.8729 ± 0.0076	0.8044 ± 0.0140	0.6646 ± 0.0100	
	4	0.8738 ± 0.0068	0.8466 ± 0.0073	0.7936 ± 0.0137	0.6683 ± 0.0110	
	5	0.8679 ± 0.0067	0.8426 ± 0.0070	0.7640 ± 0.0150	0.6434 ± 0.0103	

The AUC values followed a similar pattern to the F1 scores. CDFS-KNN and MVDFS-KNN imputation reached the highest AUC values at each missing variable level, followed by KNN imputation and MI. At five missing variables, the lowest drop in AUC with the value of 3.7% was achieved by CDFS-KNN imputation. MVDFS-KNN imputation, KNN imputation, and MI resulted in drop in AUC values by 6.8%, 15.6%, and 25.4%, respectively.

The sensitivity and specificity had a different pattern for each imputation method as the number of missing variables increased. CDFS-KNN and MVDFS-KNN imputation maintained sensitivity values closest to that of the original dataset. The sensitivity values for both methods increased slightly, then dropped and remained close to the original value, in which the drop was between 0.5% and 0.9% from the original value. For KNN imputation and MI, the sensitivity increased as the number of missing variables increased. At five missing variables, the sensitivity of KNN imputation and MI were 7.6% and 18.6% higher than the original sensitivity values, respectively. The model classified more observations in the positive class, which resulted in increasing the sensitivity but also reducing the specificity.

The specificity values dropped for all imputation methods as the number of missing variables increased. CDFS-KNN and MVDFS-KNN imputation maintained the specificity values closest to that of the original dataset. At five missing variables; the drop in specificity was 4.8% and 9.2% for CDFS-KNN and MVDFS-KNN imputation, respectively. For KNN imputation, the specificity dropped by 36.8% at five missing variables. MI was the imputation method that resulted with the largest drop in specificity (89.8%) at five variables.

The increase in sensitivity and decrease in specificity when KNN imputation and MI were used could be the result of misclassifying observation as high risk. These methods may have caused significant changes to variable values, which may have caused the model to classify more observations as high risk. This results in increasing the sensitivity and the false positive rate, and therefore, reducing the specificity.

When one variable was missing, the F1 score was highest for CDFS-KNN imputation, followed by MVDFS-KNN imputation, KNN imputation, and MI, respectively. This order remained the same with a higher number of missing variables. At five missing variables, CDFS-KNN imputation resulted in 3.3% drop in the F1 score from original values. MVDFS-KNN imputation, KNN imputation, and MI, resulted in 6.1%, 14.8%, and 28.3% drop in the F1 score, respectively. Figure 4.3 illustrates the F1 score values from one to five missing variables for each imputation method and the original complete dataset.

As the number of missing variables increased, the F1 scores for all imputation methods dropped significantly from the original value. However, the CDFS-KNN and MVDFS-KNN imputation methods maintained the model’s classification performance closest to its performance on the original dataset at all levels of missing variables.

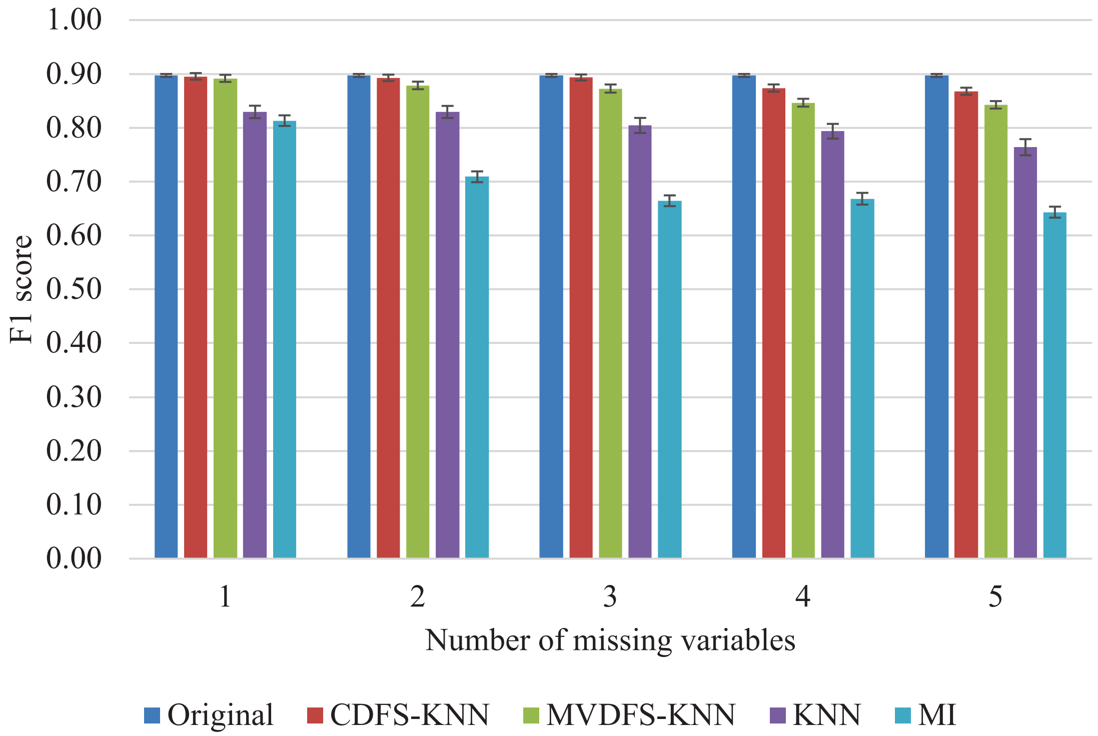


Figure 4.3: Average F1 scores for TS original and imputed datasets vs. number of missing variables.

Statistical comparisons of F1 scores at each level of missing variables are provided in Table 4.8. A two-sample t-test was conducted to compare the F1 scores between the original dataset and each dataset for different number of imputed missing variables. The level of significance was set at 0.05 for the two-sample t-test.

Table 4.8: p-values of two-sample t-tests to compare between F1 scores – TS dataset

# Missing Variables	Imputation Method	Original	CDFS-KNN	MVDFS-KNN	KNN	MI
1	Original		0.3287	0.0458	0.0000	0.0000
	CDFS-KNN	0.3287		0.4221	0.0000	0.0000
	MVDFS-KNN	0.0458	0.4221		0.0000	0.0000
	KNN	0.0000	0.0000	0.0000		0.0393
	MI	0.0000	0.0000	0.0000	0.0393	
2	Original		0.0186	0.0000	0.0000	0.0000
	CDFS-KNN	0.0186		0.0035	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0035		0.0000	0.0000

	KNN	0.0000	0.0000	0.0000	■	0.0000
	MI	0.0000	0.0000	0.0000	0.0000	■
3	**Original**	■	0.0414	0.0000	0.0000	0.0000
	CDFS-KNN	0.0414	■	0.0000	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0000	■	0.0000	0.0000
	KNN	0.0000	0.0000	0.0000	■	0.0000
	MI	0.0000	0.0000	0.0000	0.0000	■
4	**Original**	■	0.2183	0.0000	0.0000	0.0000
	CDFS-KNN	0.2183	■	0.0000	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0000	■	0.0000	0.0000
	KNN	0.0000	0.0000	0.0000	■	0.0000
	MI	0.0000	0.0000	0.0000	0.0000	■
5	**Original**	■	0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000	■	0.0000	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0000	■	0.0000	0.0000
	KNN	0.0000	0.0000	0.0000	■	0.0000
	MI	0.0000	0.0000	0.0000	0.0000	■

The F1 scores remained close to original without a statistically significant difference when one variable was missing and imputed using CDFS-KNN or MVDFS-KNN imputation. The same result was achieved by CDFS-KNN imputation when three variables were missing. KNN imputation and MI resulted in a significant drop in the F1 score at all numbers of missing variables. MVDFS-KNN imputation resulted in a significantly smaller F1 scores when two or more variables were missing.

4.2.2 Breast Cancer Wisconsin Data Experiments

For the BCW dataset, an SVM model was fitted to predict breast cancer diagnosis using 5-fold cross validation with 50 iterations. SVM was selected because multiple studies in the literature reported successful results using this algorithm on the BCW dataset to predict breast cancer diagnosis (Zheng et al., 2014, Islam et al., 2017, and Wang et al., 2018).

The top selected variables, using wrapper feature selection, which are ordered in descending importance were the following: perimeter_worst, radius_worst, area_worst, concave.points_worst, concave.points_mean, perimeter_mean, area_mean, concavity_mean, radius_mean, and area_se. The performance of the SVM model was measured using the same validation set (20% of dataset; as in the case of the TS dataset) as a reference for imputed datasets. Table 4.9 summarizes the average values of performance measures vs. number of missing variables for each imputation method. Figure 4.4 provides a visual of the F1 score drop as the number of missing variables increased.

Table 4.9: Average performance measures of SVM for BCW original and imputed datasets

Measure	# of Missing Variables	Imputation Methods				Original Dataset
		CDFS-KNN	**MVDFS-KNN**	**KNN**	**MI**	
AUC	1	0.9751 ± 0.0040	0.9734 ± 0.0035	0.9737 ± 0.0039	0.9742 ± 0.0040	0.9750 ± 0.0018
	2	0.9744 ± 0.0040	0.9744 ± 0.0035	0.9737 ± 0.0041	0.9714 ± 0.0044	
	3	0.9739 ± 0.0042	0.9741 ± 0.0037	0.9730 ± 0.0042	0.9691 ± 0.0046	
	4	0.9738 ± 0.0035	0.9704 ± 0.0035	0.9721 ± 0.0042	0.9681 ± 0.0041	
	5	0.9723 ± 0.0038	0.9689 ± 0.0038	0.9714 ± 0.0043	0.9647 ± 0.0044	
Sensitivity	1	0.9587 ± 0.0087	0.9573 ± 0.0081	0.9545 ± 0.0091	0.9569 ± 0.0092	0.9571 ± 0.0038
	2	0.9578 ± 0.0091	0.9586 ± 0.0080	0.9500 ± 0.0094	0.9539 ± 0.0091	
	3	0.9569 ± 0.0093	0.9567 ± 0.0080	0.9476 ± 0.0100	0.9499 ± 0.0092	
	4	0.9542 ± 0.0091	0.9536 ± 0.0080	0.9466 ± 0.0097	0.9409 ± 0.0087	
	5	0.9509 ± 0.0093	0.9479 ± 0.0077	0.9457 ± 0.0096	0.9343 ± 0.0083	
Specificity	1	0.9850 ± 0.0038	0.9837 ± 0.0041	0.9847 ± 0.0036	0.9848 ± 0.0036	0.9856 ± 0.0017
	2	0.9845 ± 0.0037	0.9842 ± 0.0042	0.9864 ± 0.0037	0.9829 ± 0.0040	
	3	0.9842 ± 0.0038	0.9845 ± 0.0042	0.9864 ± 0.0037	0.9818 ± 0.0042	
	4	0.9833 ± 0.0042	0.9803 ± 0.0040	0.9859 ± 0.0036	0.9842 ± 0.0038	
	5	0.9828 ± 0.0043	0.9808 ± 0.0042	0.9853 ± 0.0035	0.9821 ± 0.0044	
F1 Score	1	0.9658 ± 0.0056	0.9646 ± 0.0047	0.9633 ± 0.0056	0.9644 ± 0.0057	0.9654 ± 0.0025
	2	0.9649 ± 0.0057	0.9658 ± 0.0047	0.9624 ± 0.0057	0.9612 ± 0.0060	
	3	0.9642 ± 0.0059	0.9650 ± 0.0049	0.9611 ± 0.0060	0.9582 ± 0.0063	
	4	0.9621 ± 0.0056	0.9599 ± 0.0047	0.9600 ± 0.0059	0.9552 ± 0.0055	
	5	0.9599 ± 0.0057	0.9574 ± 0.0049	0.9561 ± 0.0060	0.9503 ± 0.0055	

The average AUC value of the SVM model decreased as the number of missing variables increased. CDFS-KNN showed a small drop (0.3%) in AUC at five missing variables. MVDFS-KNN imputation and KNN imputation caused a small drop in the AUC value as well. At four and five missing variables, the drop for MVDFS_KNN was slightly higher than the drop caused by KNN imputation. MVDFS-KNN imputation dropped the AUC value by 0.5% to 0.6%, while KNN imputation dropped it by 0.3 to 0.4%. The MI method resulted in the largest AUC drop, where it was 1.1% at five missing variables.

The sensitivity of CDFS-KNN imputation increased at one and two missing variables, at three variables it was equal to original. At four and five missing variables, it dropped by 0.4% to 0.6%. MVDFS-KNN imputation caused a similar trend in sensitivity. The drop in the AUC at 4 and five variables was between 0.4 and 1%. KNN imputation caused the sensitivity to drop as the number of missing variables increased. At five missing variables, the drop was 1.2% from the original sensitivity value. MI remained very close to the original value at one missing variable. Then it dropped continuously and reached a 2.4% drop at five missing variables.

The specificity of CDFS-KNN and MVDFS-KNN imputation dropped as the number of missing variables increased. At five variables, the drop was by 0.3% and 0.5%, respectively. For KNN imputation, the specificity increased at two and three missing variables by 0.1%, then dropped and remained closest to original value. At five variables, the KNN had the same specificity as original dataset.

The F1 score for all methods dropped as the number of missing variables increased, as shown in Figure 4.8. When one variable was missing, CDFS-KNN imputation kept the F1 score value close to original, then it started dropping at two missing variables. MCDFS-

KNN imputation dropped at one missing variables, increased at two missing variables, then started dropping at three missing variables. KNN imputation and MI caused the F1 score to drop at all numbers of missing variables. At five missing variables, all the methods resulted in a drop in the F1 score ordered in the following descending order: 0.6% for CDFS-KNN imputation, 0.8% for MVDFS-KNN imputation; 1.0% for KNN imputation; and 1.6% for MI.

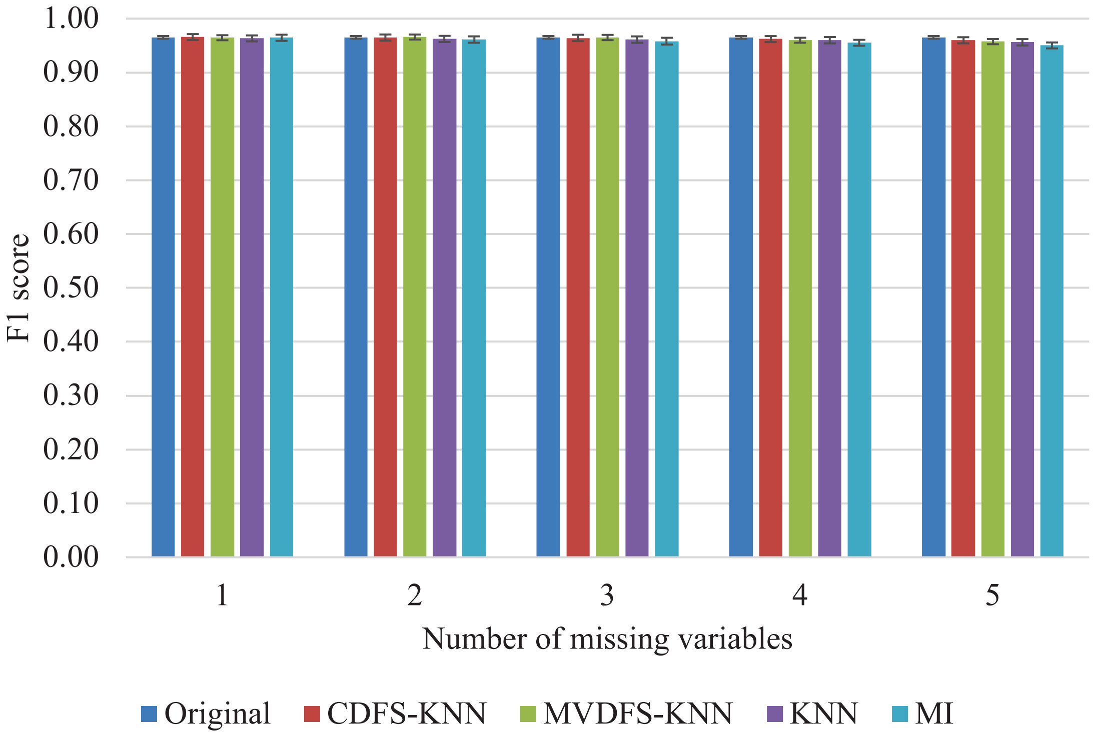


Figure 4.4: Average F1 scores for BCW original and imputed datasets vs. number of missing variables.

In general, as the number of missing variables increased, the F1 score values for all imputation methods dropped from the original value. However, the CDFS-KNN and MVDFS-KNN imputation methods maintained the model's F1 score closest the F1 score of the original dataset.

Statistical comparisons of F1 scores at each level of missing variables are provided in Table 4.10. A two-sample t-test was conducted to compare the F1 scores between the original dataset and each dataset for different number of imputed missing variables. The level of significance was set at 0.05 for the two-sample t-test.

Table 4.10: p-values of two-sample t-tests to compare between F1 scores – BCW dataset

# Missing Variables	Imputation Method	Original	CDFS-KNN	MVDFS-KNN	KNN	MI
1	**Original**		0.4986	0.8439	0.9004	0.7134
	CDFS-KNN	0.4986		0.5539	0.5371	0.7364
	MVDFS-KNN	0.8439	0.5539		0.9418	0.8255
	KNN	0.9004	0.5371	0.9418		0.7857
	MI	0.7134	0.7364	0.8255	0.7857	
2	**Original**		0.6796	0.5263	0.2863	0.1826
	CDFS-KNN	0.6796		0.8703	0.5387	0.3789
	MVDFS-KNN	0.5263	0.8703		0.6108	0.4244
	KNN	0.2863	0.5387	0.6108		0.7789
	MI	0.1826	0.3789	0.4244	0.7789	
3	**Original**		0.8929	0.8345	0.4809	0.1496
	CDFS-KNN	0.8929		0.7657	0.4721	0.1735
	MVDFS-KNN	0.8345	0.7657		0.6262	0.2344
	KNN	0.4809	0.4721	0.6262		0.5074
	MI	0.1496	0.1735	0.2344	0.5074	
4	**Original**		0.7619	0.4769	0.4423	0.0540
	CDFS-KNN	0.7619		0.6841	0.6207	0.0874
	MVDFS-KNN	0.4769	0.6841		0.8862	0.1432
	KNN	0.4423	0.6207	0.8862		0.2402
	MI	0.0540	0.0874	0.1432	0.2402	
5	**Original**		0.0067	0.0004	0.0049	0.0000
	CDFS-KNN	0.0067		0.5588	0.8634	0.0192
	MVDFS-KNN	0.0004	0.5588		0.7012	0.0344
	KNN	0.0049	0.8634	0.7012		0.0508
	MI	0.0000	0.0192	0.0508	0.0344	

When the number of missing variables was between one and four, all the imputation methods resulted in F1 scores that were not significantly different from the original F1 score. When five variables were missing, all imputation methods resulted in a significant difference from the original value; CFDS-KNN, MVDFS-KNN and KNN imputation resulted in similar F1 scores, and MI resulted in a significantly less F1 score than the rest of the imputation methods.

4.2.3 Cervical Cancer Data Experiments

An RF model was utilized to predict cervical cancer diagnosis using the CC dataset. Five-fold cross validation was used, and the experiments were repeated 50 times. Similar to TS and BCW datasets, 20% of the dataset was used for validation and the top one to five variables were simulated to be missing. The RF algorithm was selected because it was successfully applied on the CC dataset in the literature by many research studies (Abdoh et al., 2018, Alsmariy et al., 2020, and Razali et al., 2020).

The top selected variables, which used wrapper feature selection and ordered by descending importance were as follows: STDs_number, STDs, STDs_Number_of_diagnosis, Hormonal_Contraceptives_years, IUD_years, Dx, IUD, STDs_vulvo_perineal_condylomatosis, Dx_Cancer, and First_sexual_intercourse. The data is imbalanced as 12% of the records have a positive class. The training data was balanced using SMOTE. The performance of the RF model was measured for the validation set and was used as a reference for imputed datasets. Table 4.11 summarizes the average performance measures of RF for CC original and imputed datasets. Figure 4.5 provides the average F1 scores for CC original and imputed datasets vs. number of missing variables.

Table 4.11: Average performance measures of RF for CC original and imputed datasets

Measure	# of Missing Variables	Imputation Methods				Original Dataset
		CDFS-KNN	MVDFS-KNN	KNN	MI	
AUC	1	0.9322 ± 0.0039	0.9328 ± 0.0036	0.9316 ± 0.0036	0.9325 ± 0.0035	0.9326 ± 0.0016
	2	0.9287 ± 0.0035	0.9300 ± 0.0034	0.9287 ± 0.0036	0.9036 ± 0.0047	
	3	0.9205 ± 0.0035	0.9044 ± 0.0048	0.9169 ± 0.0038	0.7254 ± 0.0278	
	4	0.9172 ± 0.0032	0.9008 ± 0.0049	0.8987 ± 0.0042	0.7311 ± 0.0187	
	5	0.9111 ± 0.0034	0.8962 ± 0.0047	0.8925 ± 0.0035	0.7303 ± 0.0187	
Sensitivity	1	0.8582 ± 0.0073	0.8655 ± 0.0069	0.8573 ± 0.0074	0.8769 ± 0.0066	0.8563 ± 0.0033
	2	0.8590 ± 0.0074	0.8885 ± 0.0066	0.8595 ± 0.0075	0.9093 ± 0.0058	
	3	0.8420 ± 0.0072	0.9033 ± 0.0063	0.8359 ± 0.0078	0.9871 ± 0.0091	
	4	0.8312 ± 0.0071	0.9007 ± 0.0067	0.7973 ± 0.0123	0.9873 ± 0.0091	
	5	0.8174 ± 0.0083	0.8935 ± 0.0070	0.7771 ± 0.0102	0.9873 ± 0.0091	
Specificity	1	0.9776 ± 0.0030	0.9749 ± 0.0030	0.9771 ± 0.0029	0.9690 ± 0.0032	0.9790 ± 0.0013
	2	0.9725 ± 0.0029	0.9587 ± 0.0041	0.9720 ± 0.0031	0.9023 ± 0.0071	
	3	0.9692 ± 0.0034	0.9083 ± 0.0063	0.9671 ± 0.0036	0.1303 ± 0.0854	
	4	0.9695 ± 0.0033	0.9040 ± 0.0064	0.9597 ± 0.0032	0.1296 ± 0.0855	
	5	0.9676 ± 0.0035	0.9015 ± 0.0064	0.9606 ± 0.0029	0.1296 ± 0.0855	
F1 Score	1	0.9098 ± 0.0048	0.9124 ± 0.0044	0.9090 ± 0.0046	0.9153 ± 0.0042	0.9096 ± 0.0021
	2	0.9071 ± 0.0046	0.9157 ± 0.0040	0.9072 ± 0.0047	0.8949 ± 0.0051	
	3	0.8953 ± 0.0044	0.8949 ± 0.0054	0.8904 ± 0.0049	0.6158 ± 0.0061	
	4	0.8892 ± 0.0042	0.8911 ± 0.0055	0.8621 ± 0.0078	0.6151 ± 0.0061	
	5	0.8796 ± 0.0048	0.8856 ± 0.0054	0.8502 ± 0.0065	0.6151 ± 0.0061	

The average AUC value of the RF model decreased as the number of missing variables increased. The smallest drop in AUC was achieved by CDFS-KNN imputation, where the value dropped by 2.3% at five missing variables. At five missing variables, MVDFS-KNN imputation resulted in a 3.9% drop, KNN imputation resulted in a 4.3% drop, and MI resulted in the highest drop of 21.7%. CDFS-KNN and MVDFS-KNN imputation remained closest to the original AUC value.

The average sensitivity values change across imputation methods and number of missing variables. At five missing variables, CDFS-KNN resulted in a 4.5% drop from original sensitivity. For MVDSF-KNN imputation, the sensitivity values were higher than original at all levels of missing variables, the largest increase was by 5.5% when three variables were missing. At five missing variables it was 4.3% larger than original. The KNN imputation increased the sensitivity by 0.4% at two missing variables, then is started decreasing. At five missing variables, KNN imputation resulted in a 9.3% drop from the original value. MI sensitivity increased as the number of missing variables increased. The sensitivity of MI was 15.3% higher than the original at five missing variables. Using MI resulted in classifying more observations as positive class, which caused the sensitivity to increase and the specificity to drop.

The specificity of CDFS-KNN, MVDFS-KNN, and KNN imputation dropped as the number of missing variables increased. At five missing variables, the drop was by 1.2%, 7.9%, and 1.9%, respectively. MI caused largest drop in the specificity as the number of missing variables increased. At five missing variables, the specificity dropped by 86.8%. CDFS-KNN imputation preserved the specificity value closest to original values.

The F1 score for all methods dropped as the number of missing variables increased. At five missing variables, all the methods resulted in a drop in the F1 score ordered in the following descending order: 2.6% for MVDFS_KNN imputation; 3.3% for CDFS-KNN imputation; 6.5% for KNN imputation; and 32.4% for MI.

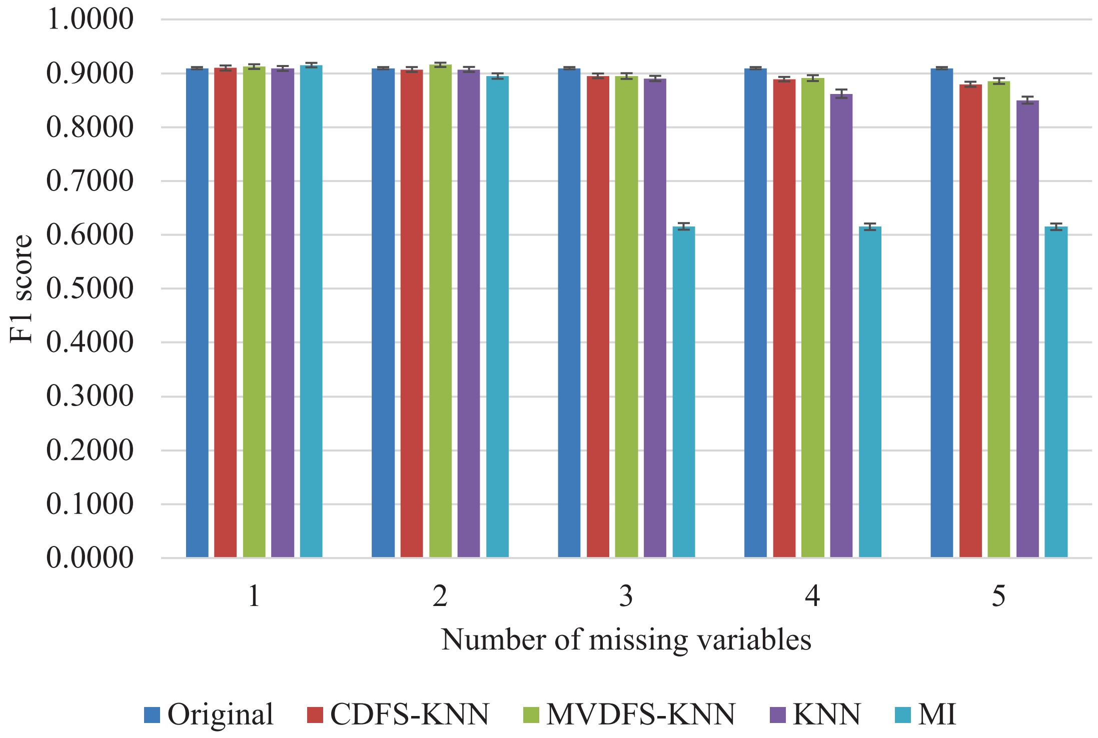


Figure 4.5: Average F1 scores for CC original and imputed datasets vs. number of missing variables.

Statistical comparisons of F1 scores at each level of missing variables are provided in Table 4.12. A two-sample t-test was conducted to compare the F1 scores between the original dataset and each dataset for different number of imputed missing variables. The level of significance was set at 0.05 for the two-sample t-test.

Table 4.12: p-values of two-sample t-tests to compare between F1 scores – CC dataset

# Missing Variables	Imputation Method	Original	CDFS-KNN	MVDFS-KNN	KNN	MI
1	**Original**	■	0.8556	0.4881	0.6459	0.0915
	CDFS-KNN	0.8556	■	0.4293	0.8057	0.0942
	MVDFS-KNN	0.4881	0.4293	■	0.2882	0.3599
	KNN	0.6459	0.8057	0.2882	■	0.0492
	MI	0.0915	0.0942	0.3599	0.0492	■
2	**Original**	■	0.8152	0.0013	0.7922	0.0006
	CDFS-KNN	0.8152	■	0.0059	0.9745	0.0007
	MVDFS-KNN	0.0013	0.0059	■	0.0076	0.0000
	KNN	0.7922	0.9745	0.0076	■	0.0007
	MI	0.0006	0.0007	0.0000	0.0007	■
3	**Original**	■	0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000	■	0.9081	0.1461	0.0000
	MVDFS-KNN	0.0000	0.9081	■	0.2275	0.0000
	KNN	0.0000	0.1461	0.2275	■	0.0000
	MI	0.0000	0.0000	0.0000	0.0000	■
4	**Original**	■	0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000	■	0.5843	0.0000	0.0000
	MVDFS-KNN	0.0000	0.5843	■	0.0000	0.0000
	KNN	0.0000	0.0000	0.0000	■	0.0000
	MI	0.0000	0.0000	0.0000	0.0000	■
5	**Original**	■	0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000	■	0.1019	0.0000	0.0000
	MVDFS-KNN	0.0000	0.1019	■	0.0000	0.0000
	KNN	0.0000	0.0000	0.0000	■	0.0000
	MI	0.0000	0.0000	0.0000	0.0000	■

When one variable was missing, all imputation methods resulted in F1 scores close to original. There was not a statistically significant difference between the scores. When two variables were missing, CDFS-KNN and KNN imputation resulted in F1 scores that were not significantly different from original. When the number of missing variables was between three and five, all the methods resulted in significantly smaller F1 scores. Additionally, CDFS-KNN and MVDFS-KNN resulted in similar F1 scores, but the other methods were significantly different.

4.2.4 Adult Data Experiments

An RF model was utilized to predict income in the Adult dataset. Five-fold cross validation was used, and the experiments were repeated 50 times. Similar to the previously discussed datasets, 20% of the dataset was used for validation where the top one to five variables were simulated to be missing. The RF algorithm was selected because it was successfully applied on this dataset in the literature by several research studies (Chakrabarty et al., 2018 and Chen et al., 2022).

The top variables, selected using wrapper feature selection, and ordered by descending importance, were capital gain, age, marital status, work hours per week, relationship, capital loss, executive managerial occupation, professional specialty occupation, and bachelor's education level. The data is somewhat balanced as 24% of the records have a positive class. The performance of the RF model was measured for the validation set and was used as a reference for imputed datasets. Table 4.13 summarizes the average performance measures of RF for Adult original and imputed datasets. Figure 4.6 provides the average F1 scores for Adult original and imputed datasets vs. number of missing variables.

Table 4.13: Average performance measures of RF for Adult original and imputed datasets

Measure	# of Missing Variables	Imputation Methods				Original Dataset
		CDFS-KNN	MVDFS-KNN	KNN	MI	
AUC	1	0.8035 ± 0.0020	0.7757 ± 0.0021	0.8030 ± 0.0023	0.8042 ± 0.0020	0.8285 ± 0.0009
	2	0.7884 ± 0.0020	0.7641 ± 0.0020	0.7883 ± 0.0021	0.7875 ± 0.0021	
	3	0.7888 ± 0.0020	0.7570 ± 0.0021	0.7844 ± 0.0023	0.8141 ± 0.0024	
	4	0.7827 ± 0.0019	0.7488 ± 0.0022	0.7674 ± 0.0024	0.8112 ± 0.0028	
	5	0.7807 ± 0.0019	0.7489 ± 0.0022	0.7642 ± 0.0027	0.8039 ± 0.0070	
Sensitivity	1	0.5433 ± 0.0036	0.5701 ± 0.0035	0.5423 ± 0.0039	0.5366 ± 0.0036	0.6306 ± 0.0016
	2	0.5682 ± 0.0038	0.5996 ± 0.0040	0.5630 ± 0.0042	0.5630 ± 0.0037	
	3	0.5678 ± 0.0039	0.5408 ± 0.0037	0.5126 ± 0.0044	0.2962 ± 0.0049	
	4	0.5761 ± 0.0037	0.5616 ± 0.0048	0.4866 ± 0.0038	0.2454 ± 0.0071	
	5	0.5782 ± 0.0037	0.5624 ± 0.0049	0.4840 ± 0.0041	0.0306 ± 0.0020	
Specificity	1	0.9397 ± 0.0010	0.9153 ± 0.0014	0.9395 ± 0.0011	0.9413 ± 0.0010	0.9396 ± 0.0005
	2	0.9249 ± 0.0013	0.8984 ± 0.0017	0.9258 ± 0.0014	0.9254 ± 0.0012	
	3	0.9252 ± 0.0013	0.9083 ± 0.0017	0.9338 ± 0.0014	0.9785 ± 0.0008	
	4	0.9191 ± 0.0013	0.8960 ± 0.0020	0.9288 ± 0.0014	0.9826 ± 0.0009	
	5	0.9171 ± 0.0013	0.8959 ± 0.0020	0.9274 ± 0.0014	0.9982 ± 0.0002	
F1 Score	1	0.6266 ± 0.0031	0.6204 ± 0.0029	0.6256 ± 0.0035	0.6232 ± 0.0031	0.6923 ± 0.0013
	2	0.6293 ± 0.0030	0.6244 ± 0.0028	0.6264 ± 0.0031	0.6259 ± 0.0031	
	3	0.6294 ± 0.0031	0.5909 ± 0.0028	0.5953 ± 0.0034	0.4337 ± 0.0054	
	4	0.6289 ± 0.0029	0.5941 ± 0.0033	0.5680 ± 0.0033	0.3759 ± 0.0087	
	5	0.6284 ± 0.0028	0.5946 ± 0.0033	0.5644 ± 0.0038	0.0588 ± 0.0036	

The AUC values for all imputation methods generally dropped as the number of missing variables increased. However, MI AUC increased slightly when three and four variables were missing but dropped at five missing variables by 3.0% from the original. CDFS-KNN imputation dropped the AUC value by 5.6% at five missing variables. KNN imputation reduced the AUC value by 7.8% and MVDF-KNN imputation by 9.6%.

The sensitivity for all imputation methods dropped at five missing variables. However, the largest drop was caused by MI. The MI sensitivity dropped by 95.2% from the original. This shows that evaluating model performance using the AUC measure only can be misleading, where high AUC values could be driven by the sensitivity or the specificity. At five missing variables, KNN imputation reduced the sensitivity by 23.2%, for MVDFS-KNN it dropped by 10.8%, and for CDFS-KNN it dropped by 8.3%.

The specificity for CDFS-KNN and MVDFS-KNN imputation decreased slightly as the number of missing variables increased. At five missing variables, the specificity dropped by 2.4% and 4.7%, respectively. At five missing variables, the specificity for KNN Imputation dropped by 1.3% from the original value. MI showed a different trend in the specificity as the number of missing variables increased. The specificity dropped at two missing variables, then it started increasing. At five missing variables, it reached 99.8%, which is 6.2% greater than the original value. This explains the high AUC values accomplished by MI. Using MI resulted in classifying more observations in the negative class, this is the reason behind the sensitivity drop and specificity increase.

The F1 scores for CDFS-KNN imputation were closest to original values as the number of missing variables increased. At five missing variables the smallest drop in the F1 score was achieved by CDFS-KNN imputation, where it was 9.2% less than the original, followed by MVDFS_KNN imputation, KNN imputation, and MI. The drop in the F1 scores was 14.1%, 18.5%, and 91.5%, respectively.

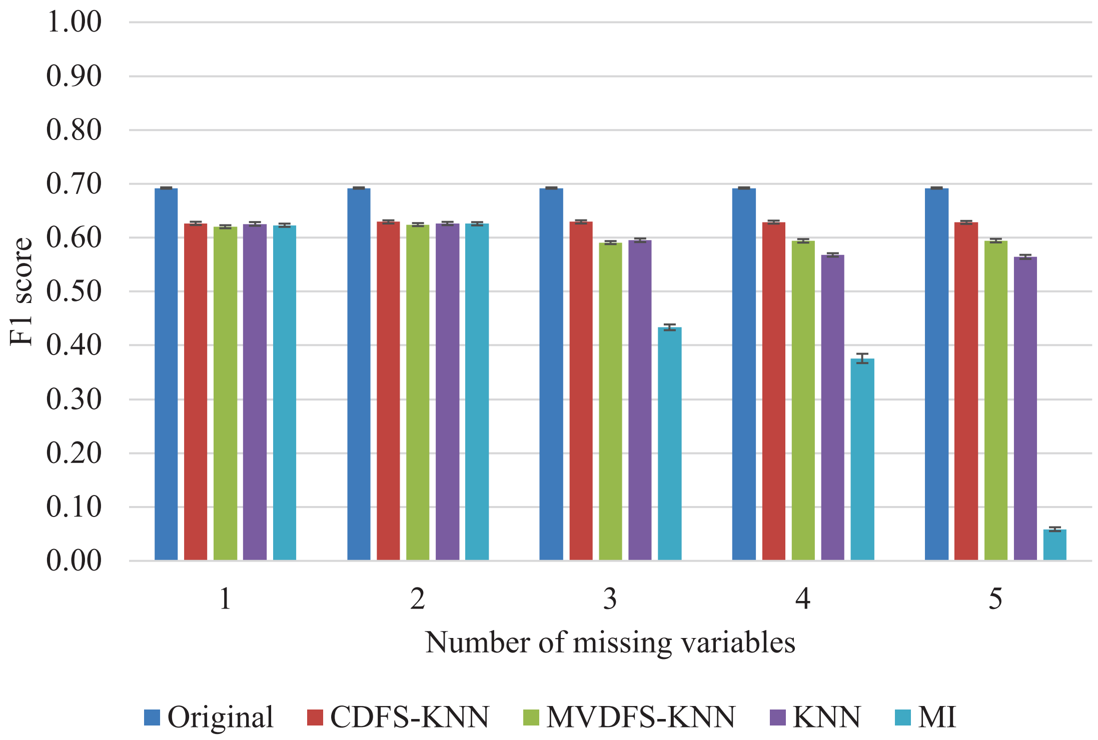


Figure 4.6: Average F1 scores for Adult original and imputed datasets vs. number of missing variables.

Statistical comparisons of F1 scores at each level of missing variables are provided in Table 4.14. A two-sample t-test was conducted to compare the F1 scores between the original dataset and each dataset for different number of imputed missing variables. The level of significance was set at 0.05 for the two-sample t-test.

Table 4.14: p-values of two-sample t-tests to compare between F1 scores – Adult dataset

# Missing Variables	Imputation Method	Original	CDFS-KNN	MVDFS-KNN	KNN	MI
1	Original		0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000		0.0039	0.6846	0.1206
	MVDFS-KNN	0.0000	0.0039		0.0222	0.1915

	KNN	0.0000	0.6846	0.0222		0.9295
	MI	0.0000	0.1206	0.1915	0.9295	
2	**Original**		0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000		0.0175	0.1739	0.1138
	MVDFS-KNN	0.0000	0.0175		0.3398	0.4589
	KNN	0.0000	0.1739	0.3398		0.8324
	MI	0.0000	0.1138	0.4589	0.8324	
3	**Original**		0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000		0.0000	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0000		0.0526	0.0000
	KNN	0.0000	0.0000	0.0526		0.0000
	MI	0.0000	0.0000	0.0000	0.0000	
4	**Original**		0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000		0.0000	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0000		0.0000	0.0000
	KNN	0.0000	0.0000	0.0000		0.0000
	MI	0.0000	0.0000	0.0000	0.0000	
5	**Original**		0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000		0.0000	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0000		0.0000	0.0000
	KNN	0.0000	0.0000	0.0000		0.0000
	MI	0.0000	0.0000	0.0000	0.0000	

At all levels of missing variables, the F1 score dropped significantly from original for all imputation methods. Some methods resulted in similar F1 scores without a significant difference such as MI and the rest of the imputation methods when one or two variables were missing. KNN and MVDFS-KNN imputation also resulted in similar F1 scores at three missing variables. However, as the number of missing variables increased to four and five, all methods had significantly different F1 scores, and the CDFS-KNN F1 score remained closest to the original.

4.2.5 Diabetes Data Experiments

An RF model was utilized to predict 30-day readmission risk in diabetes patients using the Diabetes 130 US hospitals dataset. Five-fold cross validation was used, and the experiments were repeated for 50 iterations. Similar to previously discussed datasets, 20% of the dataset was used for validation where the top one to five variables were simulated to be missing. The RF algorithm was selected because it showed outperforming accuracy to other classification methods when applied on this dataset in the literature (Rajput and Alashetty, 2022).

The top selected variables, using wrapper feature selection, ordered by descending importance were as follows: number of inpatient visits, number of emergency visits, count of diagnosis, length of stay, count of medications, discharge_disposition_id_22, discharge_disposition_id_3, discharge_disposition_id_5, discharge_disposition_id_28, discharge_disposition_id_15, number of lab procedures, number of procedures, and flag for taking diabetes medications. The data is imbalanced as 11% of the records have a positive class (readmission in 30 days from discharge). SMOTE was used to balance the dataset. The performance of the RF model was measured for the validation set and was used as a reference for imputed datasets. Table 4.15 summarizes the average performance measures of RF for Diabetes original and imputed datasets. Figure 4.7 shows the average F1 scores for Diabetes original and imputed datasets vs. number of missing variables.

Table 4.15: Average performance measures of RF for Diabetes original and imputed datasets

Measure	# of Missing Variables	Imputation Methods				Original Dataset
		CDFS-KNN	MVDFS-KNN	KNN	MI	
AUC	1	0.8221 ± 0.0009	0.7131 ± 0.0011	0.6898 ± 0.0097	0.7033 ± 0.0031	0.8856 ± 0.0003
	2	0.8078 ± 0.0009	0.7072 ± 0.0010	0.6554 ± 0.0061	0.7019 ± 0.0031	
	3	0.7802 ± 0.0009	0.7050 ± 0.0011	0.6491 ± 0.0060	0.6994 ± 0.0033	
	4	0.7517 ± 0.0009	0.6956 ± 0.0012	0.6362 ± 0.0051	0.6995 ± 0.0032	
	5	0.7426 ± 0.0009	0.6909 ± 0.0012	0.6307 ± 0.0048	0.6992 ± 0.0032	
Sensitivity	1	0.6875 ± 0.0014	0.9060 ± 0.0011	0.8130 ± 0.0238	0.9999 ± 0.0001	0.7062 ± 0.0006
	2	0.6741 ± 0.0015	0.9316 ± 0.0010	0.8835 ± 0.0234	0.9998 ± 0.0001	
	3	0.6311 ± 0.0015	0.9499 ± 0.0008	0.9100 ± 0.0185	0.9998 ± 0.0001	
	4	0.5856 ± 0.0015	0.9657 ± 0.0008	0.9254 ± 0.0162	0.9998 ± 0.0001	
	5	0.5716 ± 0.0016	0.9691 ± 0.0008	0.9295 ± 0.0156	0.9998 ± 0.0001	
Specificity	1	0.9085 ± 0.0009	0.4626 ± 0.0019	0.5284 ± 0.0527	0.0059 ± 0.0005	0.9753 ± 0.0003
	2	0.8963 ± 0.0010	0.3852 ± 0.0018	0.3290 ± 0.0546	0.0059 ± 0.0004	
	3	0.8811 ± 0.0010	0.3251 ± 0.0018	0.2706 ± 0.0495	0.0060 ± 0.0004	
	4	0.8668 ± 0.0012	0.2401 ± 0.0018	0.2149 ± 0.0419	0.0059 ± 0.0004	
	5	0.8622 ± 0.0011	0.2126 ± 0.0017	0.1969 ± 0.0392	0.0059 ± 0.0005	
F1 Score	1	0.7599 ± 0.0011	0.6909 ± 0.0011	0.6690 ± 0.0074	0.6013 ± 0.0009	0.8121 ± 0.0004
	2	0.7439 ± 0.0011	0.6772 ± 0.0010	0.6385 ± 0.0063	0.6017 ± 0.0009	
	3	0.7053 ± 0.0010	0.6667 ± 0.0011	0.6336 ± 0.0063	0.6019 ± 0.0009	
	4	0.6642 ± 0.0011	0.6484 ± 0.0010	0.6241 ± 0.0048	0.6017 ± 0.0008	
	5	0.6512 ± 0.0012	0.6420 ± 0.0010	0.6207 ± 0.0042	0.6017 ± 0.0008	

The AUC values for all imputation methods dropped as the number of missing variables increased. At five missing variables, CDFS-KNN imputation was closest to original, where it was 16.1% less. The next smallest drop was achieved by MI, then MVDFS-KNN imputation, and KNN imputation. The drop in AUC values were 21.0%, 22.0%, and 28.8%, respectively.

The sensitivity for all imputation methods except for CDFS-KNN imputation increased as the number of missing variables increased. At five missing variables, CDFS-KNN imputation resulted in 19.1% drop in the sensitivity. KNN imputation caused an increase in the sensitivity by 31.6%. MVDFS-KNN imputation increased the sensitivity by 37.2%. MI increased the sensitivity by 41.6%.

The specificity of all imputation methods decreased as the number of missing variables increased. At five missing variables, the closest specificity value to original was achieved by CDFS-KNN imputation, where it dropped by 11% from original. MVDFS-KNN, KNN, and MI, resulted in larger drop in the specificity; 78.2%, 79.8%, and 99.4%, respectively.

The F1 scores for all imputation methods dropped as the number of missing values increased. The F1 scores for CDFS-KNN imputation were closest to original as the number of missing variables increased. At five missing variables, the drop in the F1 score was 19.8% for CDFS-KNN imputation. MVDFS_KNN imputation, KNN imputation, and MI resulted in dropping the F1 score by 20.9%, 23.6%, and 25.9%, respectively.

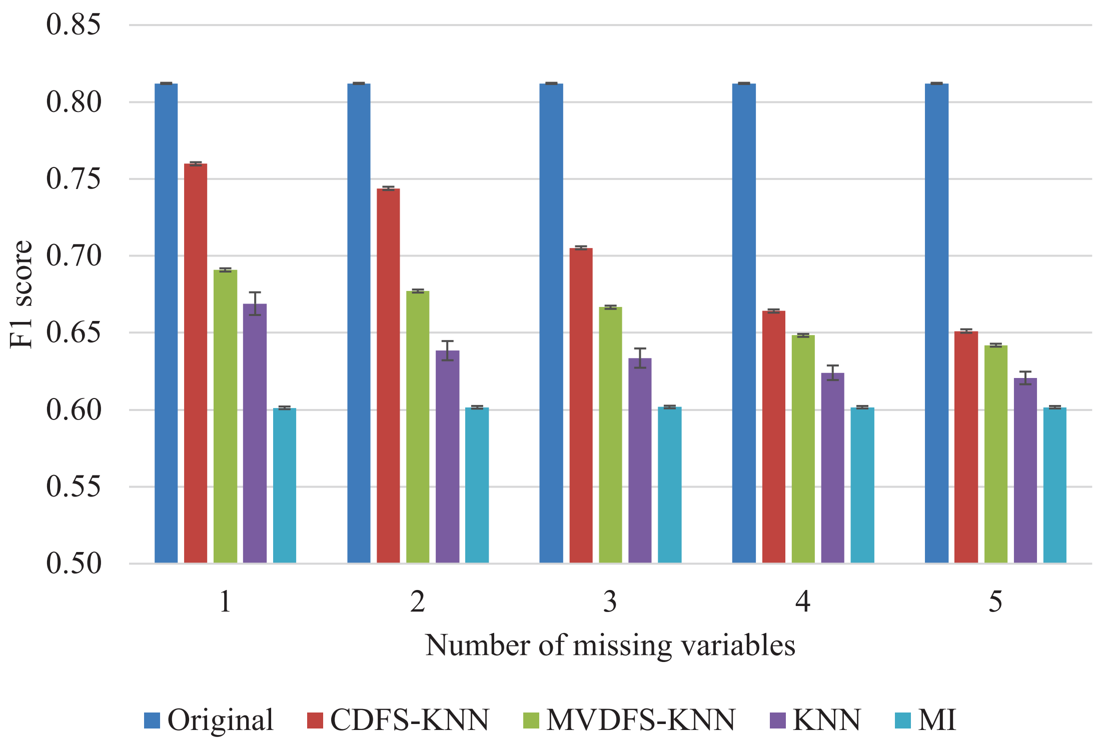


Figure 8.7: Average F1 scores for Diabetes original and imputed datasets vs. number of missing variables.

Statistical comparisons of F1 scores at each level of missing variables are provided in Table 4.16. A two-sample t-test was conducted to compare the F1 scores between the original dataset and each dataset for different number of imputed missing variables. The level of significance was set at 0.05 for the two-sample t-test.

Table 4.16: p-values of two-sample t-tests to compare between F1 scores – Diabetes dataset

# Missing Vars	Imputation Method	Original	CDFS-KNN	MVDFS-KNN	KNN	MI
1	**Original**		0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000		0.0000	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0000		0.0000	0.0000
	KNN	0.0000	0.0000	0.0000		0.0000
	MI	0.0000	0.0000	0.0000	0.0000	
2	**Original**		0.0000	0.0000	0.0000	0.0000

	CDFS-KNN	0.0000		0.0000	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0000		0.0000	0.0000
	KNN	0.0000	0.0000	0.0000		0.0000
	MI	0.0000	0.0000	0.0000	0.0000	
3	**Original**		0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000		0.0000	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0000		0.0000	0.0000
	KNN	0.0000	0.0000	0.0000		0.0000
	MI	0.0000	0.0000	0.0000	0.0000	
4	**Original**		0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000		0.0000	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0000		0.0000	0.0000
	KNN	0.0000	0.0000	0.0000		0.0000
	MI	0.0000	0.0000	0.0000	0.0000	
5	**Original**		0.0000	0.0000	0.0000	0.0000
	CDFS-KNN	0.0000		0.0000	0.0000	0.0000
	MVDFS-KNN	0.0000	0.0000		0.0000	0.0000
	KNN	0.0000	0.0000	0.0000		0.0000
	MI	0.0000	0.0000	0.0000	0.0000	

At each level of missing variables, there was a significant difference in the F1 score for all imputation methods and original dataset. Although the drop in the F1 score was significant, CDFS-KNN and MVDFS-KNN provided F1 scores closest to F1 scores generated by the original dataset.

4.3 Summary of Results

The experimental results section (section 4.2) provided detailed results of five datasets where each dataset had a different number of instances, number of features, and degree of class imbalance. For each dataset, experiments were repeated for 50 iterations to allow for statistical comparison. Data missingness was simulated for up to five important variables for the experiments. The F1 scores of all predictive models using imputation methods dropped when the number of missing variables increased. However, the results showed that CDFS-KNN imputation had the best F1 scores (closest to that of the original dataset) compared to other imputation methods. MVDFS-KNN imputation performed as second best at high levels of missingness. A comparison was conducted for all imputation methods across the five studied datasets. Figure 4.8 illustrates the average percent drop in F1 score for all imputation methods in all datasets, averaged across one to five missing variables imputed.

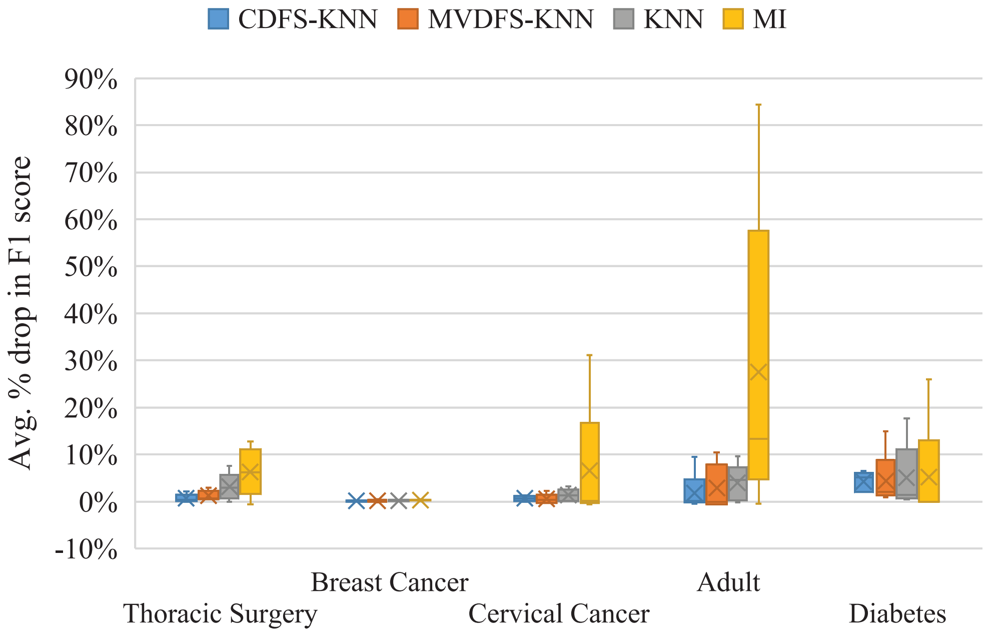
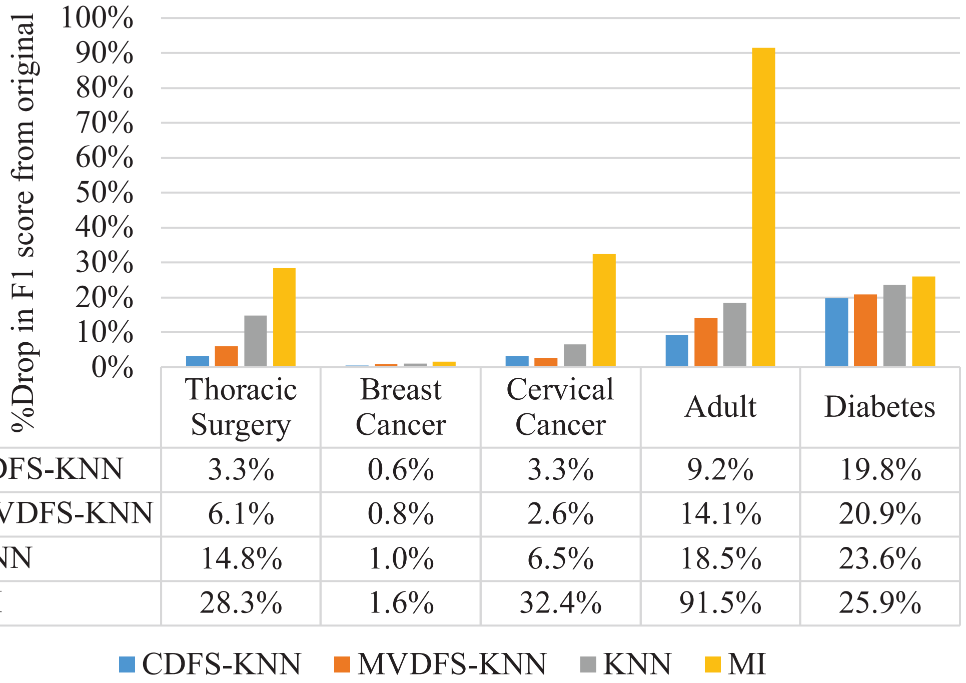


Figure 4.8: Average percent drop in F1 score averaged across one to five missing variables imputed.

MI has the largest average drop in F1 scores across all five datasets. This may be the result of changes in prospective data characteristics that may have led to false predictions. CDFS-KNN and MVDFS-KNN imputation have the least drop in F1 scores. Figure 4.9 shows the average percentage drop in F1 scores when five variables were missing compared to original F1 scores.



	Thoracic Surgery	Breast Cancer	Cervical Cancer	Adult	Diabetes
🟦 CDFS-KNN	3.3%	0.6%	3.3%	9.2%	19.8%
🟧 MVDFS-KNN	6.1%	0.8%	2.6%	14.1%	20.9%
⬛ KNN	14.8%	1.0%	6.5%	18.5%	23.6%
🟨 MI	28.3%	1.6%	32.4%	91.5%	25.9%

Figure 4.9: Average % drop in F1 score when five missing variables are imputed.

Figure 4.9 shows that MI resulted in the largest drop in F1 scores in all five datasets. The drop in F1 score was between 1.6% and 91.5% when five variables were missing. CDFS-KNN imputation had the smallest drop in F1 scores in all datasets except CC. The drop in F1 scores in CDFS-KNN imputation was between 0.6% and 19.8%. For CC the smallest drop in F1 scores was achieved by MVDFS-KNN imputation. The drop in F1 scores for MVDFS-KNN imputation was between 0.8% and 20.9%. KNN imputation had a slightly wider range for the drop in F1 scores; it was between 1.0% and 23.6%.

Further analyses were conducted to explore how the imputation methods change the characteristics of variables after imputation. The correlation between imputed and original values of variables was calculated for each imputation method and each missing variable. The cervical cancer dataset was utilized to conduct this analysis, which was repeated for 50 iterations using the 20% validation dataset. The cervical cancer dataset is considered a medium size dataset with medium dimensionality based in the scope of datasets in this research. Table 4.17 shows the average correlation values for each imputation method. Figure 4.10 summarizes the average correlation coefficient between original and imputed variables by MI. Figure 4.11 summarizes the average correlation coefficient between original and imputed variables by the rest of the imputation methods.

Table 4.17: Average correlation between original and imputed variables

Method	#Missing variables	Average correlation (0.95% CI)
MI	1	-0.0509 ± 0.0048
	2	-0.0520 ± 0.0051
	3	-0.0514 ± 0.0041
	4	-0.0485 ± 0.0036
	5	-0.0451 ± 0.0050
KNN	1	0.9596 ± 0.0010
	2	0.9625 ± 0.0005
	3	0.9118 ± 0.0009
	4	0.7540 ± 0.0027
	5	0.8781 ± 0.0014
CDFS-KNN	1	0.9085 ± 0.0013
	2	0.9675 ± 0.0005
	3	0.9422 ± 0.0006
	4	0.6288 ± 0.0028
	5	0.8965 ± 0.0018
MVDFS-KNN	1	0.9378 ± 0.0010
	2	0.9974 ± 0.0001
	3	0.9579 ± 0.0010
	4	0.7586 ± 0.0017
	5	0.8731 ± 0.0015

Table 4.17 shows that MI had the lowest correlation between imputed and original values. This is expected as MI imputed all missing values with the variable mean calculated

from the training dataset. The correlation coefficients for KNN, CDFS-KNN, and MVDFS-KNN imputation were higher compared to MI.

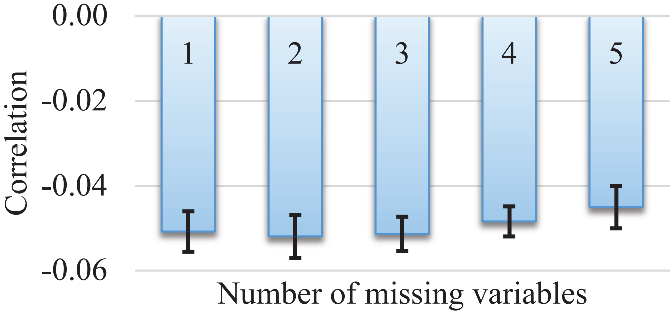


Figure 4.10: Average correlation coefficient between original and imputed variables by MI

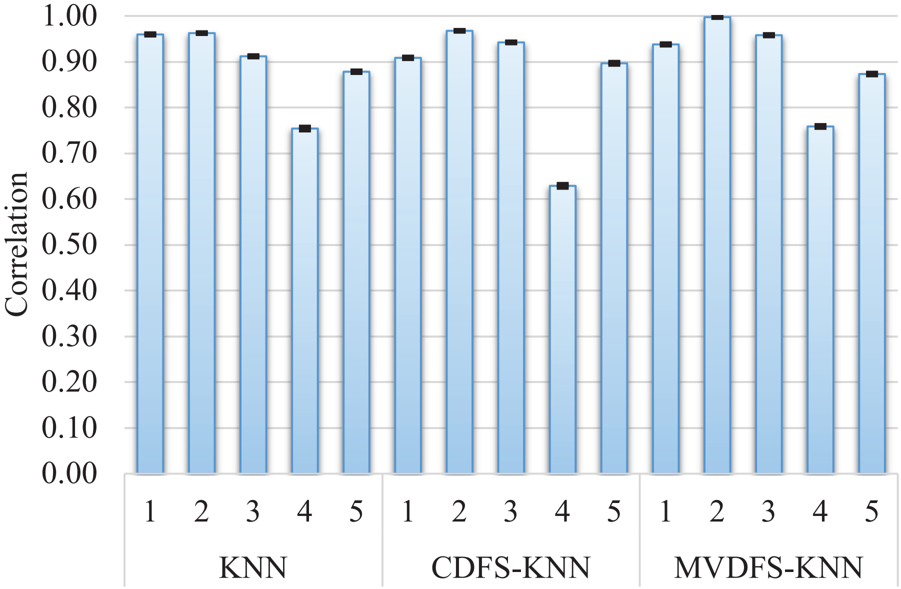


Figure 4.11: Average correlation coefficient between original and imputed variables by KNN imputation methods

The KNN imputation methods resulted in high positive correlation. This shows that these methods did not impute the missing observations with values far from the original. Because KNN methods search for similar neighbors in the training set, it is usually expected for imputed values to be close to original values. It is important for imputation methods to maintain models' performance close to original, as well as the statistical characteristics of variables after imputation.

Chapter 5: Conclusions and Future Work

This chapter provides a summary of the research area upon which this dissertation focused, the research approach, and conclusions of results. Also, it provides a discussion of future research opportunities.

5.1 Summary and Conclusions

There is a lack of implementation of predictive models in healthcare to make predictions available in a timely manner at the point of care. Missing variables in prospective data act as a barrier for deploying predictive models, where models' inputs are not complete, and thus models might not produce accurate predictions. Even though the literature is rich with predictive analytical models in healthcare, there is a limited number of publications that discuss the prospective validation and deployment of proposed predictive models in healthcare, and how models are successful in driving healthcare improvement.

This research highlighted the impact of missing prospective data on predictive models' deployment and performance. It explored utilizing complete retrospective datasets to impute missing variables in prospective datasets. This research proposed a framework for handling missing variables in prospective datasets. The proposed framework combines feature selection methods and KNN imputation to impute prospective missing variables. The proposed framework includes two hybrid approaches: CDFS-KNN imputation, which combines wrapper feature selection and KNN imputation to impute prospectively missing variables; and MVDFS-KNN imputation, which combines filter feature selection using IG and KNN imputation to impute prospectively missing variables. The proposed approaches focused on preserving models' performance with the presence of missing variables. Results were compared with common missing data imputation methods, and experiments were applied on five datasets with varying characteristics to insure robustness and generalizability. The datasets included have varying dimensionality, different number of instances, and different degrees of class imbalance. The analysis of the proposed approaches covered the impact of the number of imputed variables (one to five) on models' performance for each imputation method. Also, it measured how far the values generated by imputation methods are from original values, and how that varied by the imputation method.

The results of this study showed that the performance of predictive models on prospective data deteriorated as the number of missing variables increased. The AUC, sensitivity, specificity, and F1 score were measured for all datasets as the number of missing variables increased across all imputation methods. Imputing missing variables using MI adversely impacted models' performance, where the average percent drop in the F1 score across one to five missing variables reached 27.6% in the Adult dataset. For KNN imputation, the average percent drop in the F1 score across one to five missing variables reached 5.00% in the Diabetes dataset. For CDFS-KNN and MVDFS-KNN, the average percent drop in F1 scores were 4.30% and 4.44% in the Diabetes dataset, respectively.

When the number of missing variables was high (five), in four out of five datasets, CDFS-KNN and MVDFS-KNN maintained the F1 score closest to original, respectively; KNN imputation and MI resulted in relatively lower F1 scores, respectively. The average correlation coefficient between original and imputed values was measured for all imputation methods. For CDFS-KNN and MVDFS-KNN, the average correlation values were between 97.4% and 97.8%. This shows that the proposed approaches imputed missing values with close values relative to the original.

The proposed approaches, CDFS-KNN and MVDFS-KNN can help in designing healthcare predictive models without restrictions on using only prospectively available variables. It would also allow for deploying models at an early point of care episodes to drive preventive actions.

5.2 Future Work

Promising results have been achieved by the CDFS-KNN and MVDFS-KNN imputation approaches on five different benchmark datasets. This research can be expanded in multiple directions. Because the proposed approaches utilized complete observations in the training set and variables that are not prospectively missing, datasets with smaller dimensionality and smaller number of instances demonstrate a challenge for these methods. Smaller datasets can be used to test the proposed approaches further. The current research covered five datasets with a binary class target variable. Multiclass datasets can be used to test the performance of the proposed approaches.

Information gain was used for determining features' importance in the filter feature selection step in MVDFS-KNN. Other measures such as correlation coefficient can be used to determine features' importance. Other feature selection methods can also be used to modify the search space for the KNN imputation algorithm. This includes hybrid and embedded feature selection. These methods could result in a different set of important variables for each missing variable, and this might change imputation results and model performance metrics.

The maximum number of missing variables included in this research was five. Future work can include higher levels of missingness. This research included complete training sets. Future research can explore training sets with missing data, and the impact of missing data percentage in training sets on imputation in prospective datasets, and the prospective model performance.

CDFS-KNN and MVDFS-KNN imputation were compared to original complete datasets, KNN imputation, and MI. Other data imputation methods, such as Multiple Imputation by Chained Equations (MICE) and model-based methods, can be included in the comparison in future research.

Lastly, this research focused on the prospective missing data in healthcare problems, where prospectively missing features act as a barrier for prospective model application. Other fields of data can be included in future research, such as social sciences, manufacturing, and marketing.

Appendix

Table A: Data type and methodology used for handling missing data in reviewed healthcare predictive analytics publications

Publication	**Data Type**	**Methodology**							
		Complete Case Analysis	**LVCF**	**Keep Missing Values**	**Statistical Single Imputation**	**Statistical Multiple Imputation**	**Machine Learning Imputation**	**MLE / EM**	**Ensemble Methods**
Suh et al., 2011	Sensor Data						✓		
Bhat et al., 2011	Healthcare public dataset			✓			✓		✓
Penny et al., 2012	EMR	✓			✓		✓		
Kaambwa et al., 2012	EMR	✓				✓	✓		
Hussein et al., 2012	EMR, Sensor Data	✓		✓					
Guha et al., 2013	Cohort	✓							
Gui et al., 2014	Sensor data				✓				
Santos et al., 2015	EMR				✓		✓		
Seffens et al., 2015	Healthcare public datasets	✓							✓
Wolfson et al., 2015	EMR					✓			
Wang et al., 2015	EMR	✓					✓		
Nouaouri et al., 2015	EMR								✓
Razzaghi et al., 2015	EMR							✓	

Table A (continued)

Publication	Data Type	Methodology							
		Complete Case Analysis	LVCF	Keep Missing Values	Statistical Single Imputation	Statistical Multiple Imputation	Machine Learning Imputation	MLE / EM	Ensemble Methods
Schminkey et al., 2016	Cohort							✓	
Kayode et al., 2016	Cohort					✓			
Schuler et al., 2016	EMR, Healthcare cost data				✓	✓			
Rivers et al., 2016	Healthcare public datasets							✓	
Qian et al., 2016	EMR						✓		
Mirkes et al., 2016	EMR	✓				✓	✓		
Salgado et al., 2016	Healthcare public datasets		✓						
Conroy et al., 2016	EMR				✓	✓			✓
Srinivasan et al., 2016	Sensor Data, Survey	✓		✓	✓	✓	✓	✓	✓
Kontopantelis et al., 2017	Simulated data	✓				✓			
Ke et al., 2017	Sensor Data, Cohort					✓	✓		
Levy et al., 2017	Healthcare public datasets			✓					

Table A (continued)

Publication	Data Type	Methodology							
		Complete Case Analysis	LVCF	Keep Missing Values	Statistical Single Imputation	Statistical Multiple Imputation	Machine Learning Imputation	MLE / EM	Ensemble Methods
Saha et al., 2017	EMR						✓		
Nancy et al., 2017	Sensor Data			✓	✓		✓		✓
Zhang et al., 2017	EMR	✓							
Chen et al., 2017	EMR					✓			
Santos et al., 2017	Healthcare public datasets				✓		✓		
Kalyankar et al., 2017	Healthcare public datasets			✓	✓				
Albayrak et al., 2017	Healthcare public datasets						✓	✓	
Mayhew et al., 2018	EMR				✓	✓	✓		
Gardner et al., 2018	Cohort	✓							
Thio et al., 2018	Healthcare public datasets						✓		
Ferreira et al., 2018	EMR						✓		
Lee et al., 2018	EMR	✓							
Cao et al., 2018	Healthcare public datasets				✓	✓	✓		

Table A (continued)

Publication	Data Type	Methodology							
		Complete Case Analysis	LVCF	Keep Missing Values	Statistical Single Imputation	Statistical Multiple Imputation	Machine Learning Imputation	MLE / EM	Ensemble Methods
Wang et al., 2018	EMR						✓		
Charih et al., 2018	Healthcare public datasets	✓					✓		
Le et al., 2018	Healthcare public datasets				✓	✓	✓	✓	
Aswathi et al., 2018	Healthcare public datasets						✓		
Eirola et al., 2018	Healthcare public datasets	✓				✓	✓	✓	
Li et al., 2018	Healthcare public datasets								✓
Hegde et al., 2019	Healthcare public datasets					✓			
Phung et al., 2019	Healthcare public datasets				✓	✓	✓		
Groenhof et al., 2019	EMR				✓				
Das et al., 2019	EMR, Simulated data								✓

Table A (continued)

Publication	**Data Type**	**Methodology**							
		Complete Case Analysis	**LVCF**	**Keep Missing Values**	**Statistical Single Imputation**	**Statistical Multiple Imputation**	**Machine Learning Imputation**	**MLE / EM**	**Ensemble Methods**
Feng et al., 2019	Sensor Data			✓	✓		✓		✓
Brand et al., 2019	Simulated data, Cohort	✓			✓	✓			
Liu et al., 2019	Healthcare public datasets		✓					✓	
Curtin et al., 2019	Cohort	✓							
Azimi et al., 2019	Sensor Data						✓	✓	
Sohrabi et al., 2019	EMR	✓							
Manashty et al., 2019	Sensor Data, Healthcare public datasets, Simulated data						✓		
Alshouiliy et al., 2019	Healthcare public datasets	✓							
Mohd et al., 2019	Cohort, EMR	✓							

Table A (continued)

Publication	Data Type	Methodology							
		Complete Case Analysis	LVCF	Keep Missing Values	Statistical Single Imputation	Statistical Multiple Imputation	Machine Learning Imputation	MLE / EM	Ensemble Methods
Tu et al., 2019	Simulated data, healthcare public dataset	✓					✓		
Silveira et al., 2019	EMR	✓							
Piri et al., 2020	EMR						✓		
Wang et al., 2020	Healthcare public datasets	✓				✓	✓		✓
Wang et al., 2020	EMR			✓		✓			
Kim et al., 2020	EMR, Healthcare public dataset				✓		✓		✓
Gupta et al., 2020	Healthcare public datasets				✓		✓		✓
Hossain et al., 2020	Sensor data			✓					
Shobha et al., 2020	Healthcare public datasets						✓	✓	✓
Anagnostou et al., 2021	Healthcare public datasets				✓	✓	✓		

Table A (continued)

Publication	**Data Type**	**Methodology**							
		Complete Case Analysis	**LVCF**	**Keep Missing Values**	**Statistical Single Imputation**	**Statistical Multiple Imputation**	**Machine Learning Imputation**	**MLE / EM**	**Ensemble Methods**
Rahim et al., 2021	Healthcare public datasets				✓				
Hasan et al., 2021	Healthcare public datasets				✓				
Earnest et al., 2021	EMR			✓					
Cesario et al., 2021	EMR		✓						
Gupta et al., 2021	Survey						✓		
Ifraz et al., 2021	Healthcare public datasets	✓							
Batra et al., 2022	EMR				✓	✓	✓		
Chalkou et al., 2021	Cohort					✓			
Luo et al., 2022	Time series					✓	✓		
Kuroda et al., 2021	EMR	✓							
Kim et al., 2021	EMR	✓	✓						
Stock et al., 2021	EMR					✓			
Garnica et al., 2021	EMR	✓		✓					
Baskozos et al., 2022	EMR + Other data sources					✓			
Rasmy et al., 2022	EMR	✓							
Leiner et al., 2021	EMR	✓							
Delora et al, 2022	EMR	✓							
Li et al, 2023	EMR								✓

References

Abdoh, S. F., Rizka, M. A., & Maghraby, F. A. (2018). Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques. *IEEE Access*, 6, 59475-59485.

Albayrak, M., Turhan, K., & Kurt, B. (2017). A missing data imputation approach using clustering and maximum likelihood estimation. *2017 Medical Technologies National Congress (TIPTEKNO).*

Alshouiliy, K., Shivanna, A., Ray, S., AlGhamdi, A., & Agrawal, D. P. (2019, October). Analysis and prediction of breast cancer using AzureML platform. In *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 0212-0218). IEEE.

Alsmariy, R., Healy, G., and Abdelhafez, H. (2020). Predicting cervical cancer using machine learning methods. *International Journal of Advanced Computer Science and Applications*, 11(7).

Amarasingham, R., Patel, P. C., Toto, K. H., Nelson, L. L., Swanson, T. S., Moore, B. J., Xie, B., Zhang, S., Alvarez, K., Ma, Y., Drazner, M. H., Kollipara, U. K., & Halm, E. A. (2013). Allocating scarce resources in real-time to reduce heart failure readmissions: a prospective, controlled study. *BMJ Quality & Safety*, 22(12), 998–1005.

Amarasingham, R., Patzer, R. E., Huesch, M. D., Nguyen, N. Q., & Xie, B. (2014). Implementing electronic health care predictive analytics: considerations and challenges. *Health Affairs,* 33(7), 1148–1154

Anagnostou, P., Tasoulis, S., Vrahatis, A. G., Georgakopoulos, S., Prina, M., Ayuso-Mateos, J. L., Bickenbach, J., Bayes-Marin, I., Caballero, F.F., Egea-Cortés, L. and García-Esquinas, E. (2021). Enhancing the human health status prediction: the athlos project. *Applied Artificial Intelligence*, 35(11), 834-856.

Andridge, R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1), 40–64.

Asuncion, A., and Newman, D. (2007). UCI machine learning repository.

Aswathi, A. K., & Antony, A. (2018). An intelligent system for thyroid disease classification and diagnosis. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1261-1264). IEEE.

Azimi, I., Pahikkala, T., Rahmani, A. M., Niela-Vilén, H., Axelin, A., & Liljeberg, P. (2019). Missing data resilient decision-making for healthcare IoT through personalization: A case study on maternal health. *Future Generation Computer Systems,* 96, 297–308.

Baskozos, G., Themistocleous, A. C., Hebert, H. L., Pascal, M., John, J., Callaghan, B. C., Laycock, H., Granovsky, Y., Crombez, G., Yarnitsky, D., Rice, A.S., Smith, B. H. & Bennett, D. L. (2022). Classification of painful or painless diabetic peripheral neuropathy and identification of the most powerful predictors using machine learning models in large cross-sectional cohorts. *BMC Medical Informatics and Decision Making*, 22(1), 1-23.

Batra, S., Khurana, R., Khan, M. Z., Boulila, W., Koubaa, A., & Srivastava, P. (2022). A pragmatic ensemble strategy for missing values imputation in health records. *Entropy*, 24(4), 533.

Batista, G. E., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied artificial intelligence*, 17(5-6), 519-533.

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and Decision Making*, 16(3), 197-208.

Bhat, V. H., Rao, P. G., Krishna, S., Shenoy, P. D., Venugopal, K. R., & Patnaik, L. M. (2011b). An efficient framework for prediction in healthcare data using soft computing techniques. In *Communications in computer and information science* (pp. 522–532).

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 34(3), 483-519.

Brand, J., van Buuren, S., le Cessie, S., & van den Hout, W. (2019). Combining multiple imputation and bootstrap in the analysis of cost-effectiveness trial data. *Statistics in Medicine*, 38(2), 210-220.

Cao, W., Wang, D., Li, J., Zhou, H., Li, L., & Li, Y. (2018). Brits: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing System*, 31, 6775-6785.

Cesario, E. O., Gumiel, Y. B., Martins, M. C. M., Dias, V. M. D. C. H., Moro, C., & Carvalho, D. R. (2021). Early identification of patients at risk of sepsis in a hospital environment. *Brazilian Archives of Biology and Technology*, 64.

Chakrabarty, N., & Biswas, S. (2018). A statistical approach to adult census income level prediction. In 2018 *International Conference on Advances in Computing, Communication Control and Networking (ICACCCN),* 207-212.

Chalkou, K., Steyerberg, E., Bossuyt, P., Subramaniam, S., Benkert, P., Kuhle, J., Disanto, G., Kappos, L., Zecca, C., Egger, M. & Salanti, G. (2021). Development, validation and clinical usefulness of a prognostic model for relapse in relapsing-remitting multiple sclerosis. *Diagnostic and Prognostic Research*, 5(1), 1-16.

Charih, F., Steeves, A., Bromwich, M., Mark, A. E., Lefrançois, R., and Green, J. R. (2018). Applications of Machine Learning Methods in Retrospective Studies on Hearing. In *2018 IEEE Life Sciences Conference (LSC),* 126-129.

Chee, M. L., Ong, M. E. H., Siddiqui, F. J., Zhang, Z., Lim, S. L., Ho, A. F. W., & Liu, N. (2021). Artificial intelligence applications for COVID-19 in intensive care and emergency settings: a systematic review. *International Journal of Environmental Research and Public Health,* 18(9), 4749

Chen, C. W., Tsai, Y. H., Chang, F. R., and Lin, W. C. (2020). Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results. *Expert Systems*, 37(5), e12553.

Chen, J., Mao, S., and Yuan, Q. (2022). Salary prediction using random forest with fundamental features. In *Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021)*, 12167, 491-498. SPIE.

Chowdhury, M. H., Islam, M. K., and Khan, S. I. (2017). Imputation of missing healthcare data. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 1-6.

Cohen, D., Lee, T. B., & Sklar, D. (2016). Precalculus. Cengage Learning.

Conroy, B., Eshelman, L., Potes, C., & Xu-Wilson, M. (2016). A dynamic ensemble approach to robust classification in the presence of missing data. *Machine Learning*, 102(3), 443-463.

Curtin, D., Dahly, D. L., van Smeden, M., O'Donnell, D. P., Doyle, D., Gallagher, P., and O'Mahony, D. (2019). Predicting 1-year mortality in older hospitalized patients: external validation of the HOMR Model. *Journal of the American Geriatrics Society*, 67(7), 1478-1483.

Das, S., and Sil, J. (2019). Managing uncertainty in imputing missing symptom value for healthcare of rural India. *Health information Science and Systems*, 7(1), 5.

Dash, M., and Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(1-4), 131-156.

Dash, S., Shakyawar, S. K., Sharma, M., & Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. *Journal of Big Data*, 6(1) 1-25.

De Silva, H., & Perera, A. S. (2016). Missing data imputation using evolutionary k-Nearest neighbor algorithm for gene expression data. In *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, 141-146.

Delora, A., Mills, A., Jacobson, D., Cornett, B., Peacock, W. F., Datta, A., & Jenks, S. P. (2022). Socioeconomic and comorbid factors affecting mortality and length of Stay in COVID-19 patients. *Cureus*, 14(10).

Devin, C. J., Bydon, M., Alvi, M. A., Kerezoudis, P., Khan, I., Sivaganesan, A., McGirt, M. J., Archer, K. R., Foley, K. T., Mummaneni, P. V., Bisson, E. F., Knightly, J. J., Shaffrey, C. I., & Asher, A. L. (2018). A predictive model and nomogram for predicting return to work at 3 months after cervical spine surgery: an analysis from the Quality Outcomes Database. *Neurosurgical Focus*, 45(5), E9.

Di Tanna, G. L., Wirtz, H., Burrows, K. L., & Globe, G. (2020). Evaluating risk prediction models for adults with heart failure: A systematic literature review. *PloS One*, 15(1), e0224135.

Duggal, R., Shukla, S., Chandra, S., Shukla, B., & Khatri, S. K. (2016). Impact of selected pre-processing techniques on prediction of risk of early readmission for diabetic patients in India. *International Journal of Diabetes in Developing Countries,* 36(4), 469–476.

Earnest, A., Palmer, C., O'Reilly, G., Burrell, M., McKie, E., Rao, S., Curtis, K. & Cameron, P. (2021). Development and validation of a risk-adjustment model for mortality and hospital length of stay for trauma patients: a prospective registry-based study in Australia. *BMJ Open,* 11(8), e050795.

Eirola, E., Akusok, A., Björk, K. M., Johnson, H., & Lendasse, A. (2018). Predicting Huntington's disease: Extreme learning machine with missing values. *Proceedings of ELM-2016* (pp. 195-206). Springer, Cham.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling,* 8(3), 430-457.

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, 8(1), 1-37.

Escobar, G. J., Puopolo, K. M., Wi, S., Turk, B. J., Kuzniewicz, M. W., Walsh, E. M., Newman, T. B., Zupancic, J. A., Lieberman, E., & Draper, D. (2014). Stratification of risk of early-onset sepsis in newborns ≥34 weeks' gestation. *Pediatrics*, 133(1), 30–36.

Feng, T., & Narayanan, S. (2019). Imputing missing data in large-scale multivariate biomedical wearable recordings using bidirectional recurrent neural networks with temporal activation regularization. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).* 2529-2534.

Fernandes, K., Cardoso, J. S., & Fernandes, J. (2017). Transfer learning with partial observability applied to cervical cancer screening. In *Lecture Notes in Computer Science* (pp. 243–250).

Ferrão, J., Oliveira, M. D., Janela, F., & Martins, H. (2016). Preprocessing structured clinical data for predictive modeling and decision support. *Applied Clinical Informatics*, 07(04), 1135–1153.

Ferreira-Santos, D., Monteiro-Soares, M., & Rodrigues, P. P. (2018). Impact of imputing missing data in Bayesian Network Structure Learning for Obstructive sleep apnea diagnosis. *PubMed*, 247, 126–130.

García-Laencina, P. J., Sancho-Gómez, J., & Figueiras-Vidal, A. R. (2009). Pattern classification with missing data: a review. Neural Computing and Applications, 19(2), 263–282.

Gardner, G., Ziauddeen, N., &Alwan, N. (2018). Investigating maternal risk factors for stillbirth in a population-based cohort in the South of England. *Revue d'Épidémiologie et de Santé Publique*, 66, S273.

Garies, S., Cummings, M., Quan, H., McBrien, K., Drummond, N., Manca, D., &Williamson, T. (2020). Methods to improve the quality of smoking records in a primary care EMR database: exploring multiple imputation and pattern-matching algorithms. *BMC Medical Informatics and Decision Making,* 20(1), 1-10.

Garnica, O., Gómez, D., Ramos, V., Hidalgo, J. I., & Ruiz-Giardín, J. M. (2021). Diagnosing hospital bacteraemia in the framework of predictive, preventive and personalised medicine using electronic health records and machine learning classifiers. *EPMA Journal*, 12(3), 365-381.

Glowacka, K. J., Henry, R. M., & May, J. H. (2009). A hybrid data mining/simulation approach for modelling outpatient no-shows in clinic scheduling. *Journal of the Operational Research Society*, 60(8), 1056–1068.

Groenhof, T. K. J., Rittersma, Z. H., Bots, M. L., Brandjes, M., Jacobs, J. J., Grobbee, D. E., Van Solinge, W. W., Visseren, F. L. J., Haitjema, S., & Asselbergs, F. W. (2019). A computerised decision support system for cardiovascular risk management 'live' in the electronic health record environment: development, validation and implementation—the Utrecht Cardiovascular Cohort Initiative. *Netherlands Heart Journal,* 27(9), 435–442.

Guha, S., Van Belle, V., Bottomley, C., Preisler, J., Vathanan, V., Sayasneh, A., Stalder, C., Timmerman, D. & Bourne, T. (2013). External validation of models and simple scoring systems to predict miscarriage in intrauterine pregnancies of uncertain viability. *Human Reproduction*, 28(11), 2905-2911.

Gui, Q., Jin, Z., & Xu, W. (2014). Exploring missing data prediction in medical monitoring: A performance analysis approach. In *2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, 1-6.

Guo, A., Mazumder, N. R., Ladner, D. P., & Foraker, R. E. (2021). Predicting mortality among patients with liver cirrhosis in electronic health records with machine learning. *PloS One,* 16(8), e0256428.

Gupta, A., Jain, V., & Singh, A. (2022). Stacking ensemble-based intelligent machine learning model for predicting post-COVID-19 complications. *New Generation Computing*, 40(4), 987-1007.

Gupta, M., & Gupta, B. (2020). A new scalable approach for missing value imputation in high-throughput microarray data on apache spark. *International Journal of Data Mining and Bioinformatics*, 23(1), 79-100.

Hasan, M. K., Jawad, M. T., Dutta, A., Awal, M. A., Islam, M. A., Masud, M., & Al-Amri, J. F. (2021). Associating measles vaccine uptake classification and its underlying factors using an ensemble of machine learning models. *IEEE Access*, 9, 119613-119628.

Hegde, H., Shimpi, N., Panny, A., Glurich, I., Christie, P., & Acharya, A. (2019). MICE vs PPCA: Missing data imputation in healthcare. *Informatics in Medicine Unlocked*, 17, 100275.

Hickey, G. L., & Blackstone, E. H. (2016). External model validation of binary clinical risk prediction models in cardiovascular and thoracic surgery. *The Journal of Thoracic and Cardiovascular Surgery,* 152(2), 351–355.

Hossain, T., Ahad, M., Rahman, A., & Inoue, S. (2020). A method for sensor-based activity recognition in missing data scenario. *Sensors*, 20(14), 3811.

Hu, Z., Melton, G. B., Arsoniadis, E. G., Wang, Y., Kwaan, M. R., & Simon, G. J. (2017). Strategies for handling missing clinical data for automated surgical site infection detection from the electronic health record. *Journal of Biomedical Informatics*, 68, 112-120.

Hussein, A. S., Omar, W. M., Li, X., & Ati, M. (2012). Efficient chronic disease diagnosis prediction and recommendation system. *In 2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences,* 209-214.

Ifraz, G. M., Rashid, M. H., Tazin, T., Bourouis, S., & Khan, M. M. (2021). Comparative analysis for prediction of kidney disease using intelligent machine learning methods. *Computational and Mathematical Methods in Medicine*, 2021. 6141470.

Islam, M. M., Iqbal, H., Haque, M. R., & Hasan, M. K. (2017). Prediction of breast cancer using support vector machine and K-Nearest neighbors. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*, 226-229.

Jabbar, M. A., B. L. Deekshatulu, & Priti Chandra. (2015). Computational intelligence technique for early diagnosis of heart disease. In *2015 IEEE International Conference on Engineering and Technology (ICETECH),* 1-6.

Jian, Y., Pasquier, M., Sagahyroon, A., & Aloul, F. (2021). A machine learning approach to predicting diabetes complications. *Healthcare*, 9(12), 1712.

Jović, A., Brkić, K., & Bogunović, N. (2015). A review of feature selection methods with applications. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1200-1205.

Kaambwa, B., Bryan, S., & Billingham, L. (2012). Do the methods used to analyse missing data really matter? An examination of data from an observational study of Intermediate Care patients. *BMC Research Notes,* 5(1), 1-12

Kalyankar, G. D., Poojara, S. R., & Dharwadkar, N. V. (2017). Predictive analysis of diabetic patient data using machine learning and Hadoop. In *2017 International Conference on I-SMAC (IoT in social, mobile, analytics and Cloud) (I-SMAC)*, 619-624.

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, 64(5), 402-406.

Kayode, G. A., Grobbee, D. E., Amoakoh-Coleman, M., Adeleke, I. T., Ansah, E., De Groot, J. A., & Klipstein-Grobusch, K. (2016). Predicting stillbirth in a low resource setting. *BMC Pregnancy and Childbirth,* 16(1), 274.

Ke, C., Jin, Y., Evans, H., Lober, B., Qian, X., Liu, J., & Huang, S. (2017). Prognostics of surgical site infections using dynamic health data. *Journal of Biomedical Informatics,* 65, 22-33.

Kim, J. C., & Chung, K. (2020). Multi-modal stacked denoising autoencoder for handling missing data in healthcare big data. *IEEE Access,* 8, 104933-104943.

Kim, K., Yang, H., Yi, J., Son, H. E., Ryu, J. Y., Kim, Y. C., Jeong, J.C., Chin, H.J., Na, K.Y., Chae, D.W. & Han, S.S. (2021). Real-time clinical decision support based on recurrent neural networks for in-hospital acute kidney injury: External validation and model interpretation. *Journal of Medical Internet Research*, 23(4), e24120.

Kohavi, R. (1996). Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. *Knowledge Discovery and Data Mining*, 202–207

Kontopantelis, E., White, I. R., Sperrin, M., & Buchan, I. (2017). Outcome-sensitive multiple imputation: a simulation study. *BMC medical research methodology*, 17(1), 2.

Kuroda, S., Matsumoto, S., Sano, T., Kitai, T., Yonetsu, T., Kohsaka, S., Torii, S., Kishi, T., Komuro, I., Hirata, K.I. & Node, K. (2021). External validation of the 4C Mortality Score for patients with COVID-19 and pre-existing cardiovascular diseases/risk factors. *BMJ Open,* 11(9), e052708.

Le, T. D., Beuran, R., & Tan, Y. (2018). Comparison of the most influential missing data imputation algorithms for healthcare. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, 247-251.

Lee, M. J., Park, J. H., Moon, Y. R., Jo, S. Y., Yoon, D., Park, R. W., Jeong, J. C., Park, I., Shin, G. T., & Kim, H. (2018). Can we predict when to start renal replacement therapy in patients with chronic kidney disease using 6 months of clinical data?. *PloS One*, 13(10), e0204586.

Leiner, J., Pellissier, V., Nitsche, A., König, S., Hohenstein, S., Nachtigall, I., Hindricks, G., Kutschker, C., Rolinski, B., Gebauer, J. & Prantz, A. (2021). SARS-CoV-2 rapid antigen testing in the healthcare sector: A clinical prediction model for identifying false negative results. *International Journal of Infectious Diseases*, 112, 117-123.

Levy, S., Duda, M., Haber, N., & Wall, D. P. (2017). Sparsifying machine learning models identify stable subsets of predictive features for behavioral detection of autism. *Molecular Autism*, 8(1), 65.

Li, B., Jin, Y., Yu, X., Song, L., Zhang, J., Sun, H., Liu, H., Shi, Y. & Kong, F. (2023). MVIRA: A model based on Missing Value Imputation and Reliability Assessment for mortality risk prediction. *International Journal of Medical Informatics*, 178, 105191.

Li, X., & Li, J. (2018). Health risk prediction using big medical data-a collaborative filtering-enhanced deep learning approach. In *2018 IEEE 20th International Conference on e-Health Networking, Applications and Services (Healthcom)*, 1-7.

Lim, S. M. S., Wong, P. L., Sulaiman, H., Atiya, N., Shunmugam, R. H., & Liew, S. M. (2019). Clinical prediction models for ESBL-Enterobacteriaceae colonization or infection: a systematic review. *Journal of Hospital Infection,* 102(1), 8-16.

Liu, M., Stella, F., Hommersom, A., Lucas, P. J., Boer, L., & Bischoff, E. (2019). A comparison between discrete and continuous time Bayesian networks in learning from clinical time series data with irregularity. *Artificial Intelligence in Medicine,* 95, 104-117.

Luo, Y. (2022). Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics*, 23(1), bbab489.

Malik, M., Abdallah, S., & Ala'raj, M. (2016). Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review. *Annals of Operations Research,* 270(1–2), 287–312.

Manashty, A., & Light, J. (2019). Life model: A novel representation of life-long temporal sequences in health predictive analytics. *Future Generation Computer Systems*, 92, 141-156.

Mayhew, M. B., Petersen, B. K., Sales, A. P., Greene, J. D., Liu, V. X., & Wasson, T. S. (2018). Flexible, cluster-based analysis of the electronic medical record of sepsis with composite mixture models. *Journal of Biomedical Informatics*, 78, 33-42.

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, 10(1), 1-16.

Mirkes, E. M., Coats, T. J., Levesley, J., & Gorban, A. N. (2016). Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. *Computers in Biology and Medicine*, 75, 203-216.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group*. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement*. Annals of Internal Medicine*, 151(4), 264-269.

Nancy, J. Y., Khanna, N. H., & Arputharaj, K. (2017). Imputing missing values in unevenly spaced clinical time series data to build an effective temporal classification framework. *Computational Statistics and Data analysis*, 112, 63-79.

Nijman, S. W. J., Leeuwenberg, A. M., Beekers, I., Verkouter, I., Jacobs, J. J. L., Bots, M. L., Asselbergs, F.W., Moons, K.G.M. & Debray, T.P.A. (2022). Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. *Journal of Clinical Epidemiology*, 142, 218-229.

Nouaouri, I., Samet, A., & Allaoui, H. (2015). Evidential data mining for length of stay (LOS) prediction problem. In *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, 1415-1420.

Nugroho, H., & Surendro, K. (2019). Missing data problem in predictive analytics. In *Proceedings of the 2019 8th International Conference on Software and Computer Applications.* 95-100.

Pang, X., Forrest, C. B., Lê-Scherban, F., & Masino, A. J. (2021). Prediction of early childhood obesity with machine learning and electronic health record data. *International Journal of Medical Informatics*, 150, 104454.

Payrovnaziri, S. N., Xing, A., Salman, S., Liu, X., Bian, J., & He, Z. (2021). Assessing the impact of imputation on the interpretations of prediction models: A case study on mortality prediction for patients with acute myocardial infarction. *AMIA Summits on Translational Science Proceedings*, 465.

Penny, K. I., & Atkinson, I. (2011). Approaches for dealing with missing data in health care studies. *Journal of Clinical Nursing*, 21(19pt20), 2722–2729.

Percac-Lima, S., Cronin, P. R., Ryan, D. P., Chabner, B. A., Daly, E. A., & Kimball, A. B. (2015). Patient navigation based on predictive modeling decreases no-show rates in cancer care. *Cancer*, 121(10), 1662–1670.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, 4(2), 1883.

Phung, S., Kumar, A., & Kim, J. (2019). A deep learning technique for imputing missing healthcare data. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6513-6516.

Piri, S. (2020). Missing care: A framework to address the issue of frequent missing values; The case of a clinical decision support system for Parkinson’s disease. *Decision Support Systems*, 136, 113339

Qian, T., & Masino, A. J. (2016). Latent patient cluster discovery for robust future forecasting and new-patient generalization*.* *Plos One*, 11(9), e0162812.

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in bioinformatics,* 2, 927312.

Rahim, A., Rasheed, Y., Azam, F., Anwar, M. W., Rahim, M. A., & Muzaffar, A. W. (2021). An integrated machine learning framework for effective prediction of cardiovascular diseases. *IEEE Access*, 9, 106575-106588.

Rajput, G. G., & Alashetty, A. (2022). A Machine learning approach to reduce the diabetes patient's readmission risk using a novel preprocessing technique. In *2022 4th International Conference on Circuits, Control, Communication and Computing (I4C)*, 173-177.

Rasmy, L., Nigo, M., Kannadath, B. S., Xie, Z., Mao, B., Patel, K., Zhou, Y., Zhang, W., Ross, A., Xu, H. & Zhi, D. (2022). Recurrent neural network models (CovRNN) for predicting outcomes of patients with COVID-19 on admission to hospital: model development and validation using electronic health record data. *The Lancet Digital Health,* 4(6), 415-42.

Ravichandran, A., Mahulikar, K., Agarwal, S., & Sankaranarayanan, S. (2021). Post thoracic surgery life expectancy prediction using machine learning. *International Journal of Healthcare Information Systems and Informatics (IJHISI),* 16(4), 1-20.

Razali, N., Mostafa, S. A., Mustapha, A., Wahab, M. H. A., & Ibrahim, N. A. (2020). Risk factors of cervical cancer using classification in data mining. *Journal of Physics*, 1529(2), 022102.

Razzaghi, T., Roderick, O., Safro, I., & Marko, N. (2016). Multilevel weighted support vector machine for classification on healthcare data with missing values. *PloS One*, 11(5), e0155119.

Rivers, C. M., Majumder, M. S., & Lofgren, E. T. (2016). Risks of death and severe disease in patients with Middle East respiratory syndrome coronavirus, 2012–2015. *American Journal of Epidemiology*, 184(6), 460-464.

Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. (2016). Big data analytics to improve cardiovascular care: promise and challenges. *Nature Reviews Cardiology*, 13(6), 350–359.

Saha, B., Gupta, S., Phung, D., & Venkatesh, S. (2017). Effective sparse imputation of patient conditions in electronic medical records for emergency risk predictions. *Knowledge and Information Systems*, 53(1), 179-206.

Salgado, C. M., Vieira, S. M., Mendonça, L. F., Finkelstein, S., & Sousa, J. M. (2016). Ensemble fuzzy models in personalized medicine: Application to vasopressors administration. *Engineering Applications of Artificial Intelligence,* 49, 141-148.

Sanchez-Pinto, L. N., Luo, Y., & Churpek, M. M. (2018). Big data and data science in critical care. *Chest*, 154(5), 1239–1248.

Santos, M. S., Abreu, P. H., García-Laencina, P. J., Simão, A., & Carvalho, A. (2015). A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of Biomedical Informatics,* 58, 49-59.

Santos, M. S., Soares, J. P., Abreu, P. H., Araújo, H., & Santos, J. (2017). Influence of data distribution in missing data imputation. In *Conference on Artificial Intelligence in Medicine in Europe,* 285-294.

Schminkey, D. L., von Oertzen, T., & Bullock, L. (2016). Handling missing data with multilevel structural equation modeling and full information maximum likelihood techniques. *Research in Nursing and Health*, 39(4), 286-297.

Schuler, A., Liu, V., Wan, J., Callahan, A., Udell, M., Stark, D. E., & Shah, N. H. (2016). Discovering patient phenotypes using generalized low rank models. *In Biocomputing 2016: Proceedings of the Pacific Symposium*, 144-155.

Seffens, W., Evans, C., & Taylor, H. A. (2015). Machine learning data imputation and classification in a multicohort hypertension clinical study. *Bioinformatics and Biology Insights*, 9s3, BBI.S29473.

Shobha, K., & Savarimuthu, N. (2020). Clustering based imputation algorithm using unsupervised neural network for enhancing the quality of healthcare data*. Journal of Ambient Intelligence and Humanized Computing,* 1-11.

Silveira, A., Muñoz, C., & Mendoza, L. (2019). Severe asthma exacerbations prediction using neural networks. In *International Conference on Engineering Applications of Neural Networks*, 115-124.

Sohrabi, B., Vanani, I. R., Gooyavar, A., & Naderi, N. (2019). Predicting the readmission of heart failure patients through data analytics. *Journal of Information and Knowledge Management*, 18(01), 1950012.

Srinivasan, K., Currim, F., Ram, S., Lindberg, C., Sternberg, E., Skeath, P., Najafi, B., Razjouyan, J., Lee, H.K., Foe-Parker, C. & Goebel, N. (2016, April). Feature importance and predictive modeling for multi-source healthcare data with missing values. In *Proceedings of the 6th International Conference on Digital Health Conference,* 47-54.

Stiglic, G., Kocbek, P., Fijacko, N., Sheikh, A., & Pajnkihar, M. (2019). Challenges associated with missing data in electronic health records: a case study of a risk prediction model for diabetes using data from Slovenian primary care. *Health Informatics Journal*, 25(3), 951-959.

Stock, S. J., Horne, M., Bruijn, M., White, H., Boyd, K. A., Heggie, R., Wotherspoon, L., Aucott, L., Morris, R.K., Dorling, J. & Jackson, L. (2021). Development and validation of a risk prediction model of preterm birth for women with preterm labour symptoms (the QUIDS study): A prospective cohort study and individual participant data meta-analysis. *PLoS Medicine*, 18(7), e1003686.

Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed Research International,* 2014, 11

Suh, M. K., Woodbridge, J., Lan, M., Bui, A., Evangelista, L. S., & Sarrafzadeh, M. (2011). Missing data imputation for remote CHF patient monitoring systems. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3184-3187.

Teo, K., Yong, C. W., Muhamad, F., Mohafez, H., Hasikin, K., Xia, K., Qian, P., Dhanalakshmi, S., Utama, N. P., & Lai, K. W. (2021). The promise for reducing healthcare cost with predictive model: an analysis with quantized evaluation metric on readmission. *Journal of Healthcare Engineering*, 2021, 1–10.

Thio, Q. C., Karhade, A. V., Ogink, P. T., Raskin, K. A., Bernstein, K. D. A., Calderon, S. A. L., & Schwab, J. H. (2018). Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma?. *Clinical Orthopedics and Related Research*, 476(10), 2040.

Trifonova, O. P., Lokhov, P. G., & Archakov, A. I. (2013). Metabolic profiling of human blood. Biochemistry (Moscow) Supplement Series B: *Biomedical Chemistry,* 7(3), 179-18

Tu, R., Zhang, C., Ackermann, P., Mohan, K., Kjellström, H., & Zhang, K. (2019). Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1762-1770.

Urbanowicz, R. J., Meeker, M., La Cava, W., Olson, R. S., & Moore, J. H. (2018). Relief-based feature selection: Introduction and review. *Journal of Biomedical Informatics*, 85, 189–203.

Wang, D. D., Ng, S. H. X., Abdul, S. N. B., Ramachandran, S., Sridharan, S., & Tan, X. Q. (2018). Imputation of missing diagnosis of diabetes in an administrative EMR system. In *2018 11th Biomedical Engineering International Conference (BMEiCON),* 1-5.

Wang, G., Lu, J., Choi, K. S., & Zhang, G. (2018). A transfer-based additive LS-SVM classifier for handling missing data. *IEEE transactions on Cybernetics,* 50(2), 739-752.

Wang, H., Huang, Z., Zhang, D., Arief, J., Lyu, T., & Tian, J. (2020). Integrating co-clustering and interpretable machine learning for the prediction of intravenous immunoglobulin resistance in kawasaki disease. *IEEE Access,* 8, 97064-97071.

Wang, H., Yao, Y. D., Qian, W., & Lure, F. (2015). Applying matrix factorization in data reconstruction for heart disease patient classification. In *2015 17th International Conference on E-health Networking, Application and Services (HealthCom)*, 245-249.

Wang, H., Zheng, B., Yoon, S. W., & Ko, H. S. (2018). A support vector machine-based ensemble algorithm for breast cancer diagnosis. *European Journal of Operational Research,* 267(2), 687-699

Wells, B. J., Chagin, K. M., Nowacki, A. S., & Kattan, M. W. (2013). Strategies for handling missing data in electronic health record derived data. *The Journal for Electronic Health Data and Methods,* 1(3): 7

Wolberg,W., Mangasarian,O., Street,N., & Street,W. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository.

Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., Mallett, S., & PROBAST Group†. (2019). PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine*, 170(1), 51-58.

Wolfson, J., Bandyopadhyay, S., Elidrisi, M., Vazquez-Benitez, G., Vock, D. M., Musgrove, D., D., Adomavicius, G., Johnson, P.E. & O'Connor, P.J. (2015). A Naive Bayes machine learning approach to risk prediction using censored, time-to-event data. *Statistics in Medicine*, 34(21), 2941-2957.

Yadav, M. L., & Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. *Knowledge-Based Systems*, 160, 104-118.

Yang, X., Tong, Y., Meng, X., Zhao, S., Xu, Z., Li, Y., Jia, X. & Tan, S. (2016). Adaptive logistic group Lasso method for predicting the no-reflow among the multiple types of high-dimensional variables with missing data. In 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), 1085-1089.

Zhang, Z., & Hong, Y. (2017). Development of a novel score for the prediction of hospital mortality in patients with severe sepsis: the use of electronic healthcare records with LASSO regression. *Oncotarget*, 8(30), 49637.

Zheng, B., Yoon, S. W., & Lam, S. S. (2014). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications*, 41(4), 1476-1482.

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods,* 12(10), 931-934.

Zięba, M., Tomczak, J. M., Lubicz, M., & Świątek, J. (2014). Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied Soft Computing,* 14, 99-108.