Journal of Theoretical and Applied Information Technology
30th April 2024. Vol.102. No 8
© Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

INTELLIGENT IMPUTATION OF MISSING DATA USING BIDIRECTIONAL NEIGHBOR GRAPH MODELING FOR DIABETIC RISK PREDICTION

BASHAR HAMAD AUBAIDAN[1], RABIAH ABDUL KADIR[1*], MOHAMAD TAHA LJAB[1], BAKR AHMED TAHA[2]

[1]*Institute of visual informatics, universiti kebangsaan malayia, bangi 43500 malaysia*

[2]*Photonics Technology Lab, Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering and Built Environment, University Kebangsaan Malaysia, UKM Bangi 43600, Malaysia*

E-mail: P103708@siswa.ukm.edu.my; rabiahivi@ukm.edu.my ; taha@ukm.edu.my ; bakerkunfu955@gmail.com.

ABSTRACT

Diabetes is a global health issue, affecting countless individuals. This study dives into the critical task of filling gaps in data crucial for diabetes forecasts. These gaps can weaken the reliability of medical datasets, leading to less effective diagnostics and predictions. We put forward a cutting-edge approach employing the bidirectional neighbour graph (BNG) algorithm. This graph-based, semi-supervised learning technique is adept at managing the intricacies of incomplete data. Compared to traditional machine learning methods, the BNG algorithm shines by forming a network where nodes represent patients. Each node links to its closest neighbors in both directions, ensuring thorough data assimilation. Our method stands out, showcasing an Area Under the Curve (AUC) score of 0.86. This score reflects a strong model with the ability to extract extensive and distinct features from the data, resulting in enhanced classification accuracy. The study highlights the necessity of precisely tackling missing values to boost model trustworthiness. Moreover, it proposes extending the BNG technique to various sectors where maintaining data accuracy is crucial. Thanks to its computational efficiency and adaptability, the BNG algorithm is championed as a flexible instrument for medical researchers. It sets the stage for more precise and dependable diabetes risk assessment and management.

Keywords: *Graph-Based Methods, Bidirectional Neighbour Graph, Classification, Feature Extraction, Machine Learning.*

1. INTRODUCTION

The field of medical data analysis is constantly evolving, driven by the increasing complexity of medical data and the need for more accurate and reliable analysis. With advances in medical technology and the availability of large amounts of healthcare data, there is a growing demand for innovative solutions that can effectively handle the challenges posed by complex medical data [1]. The motivation behind the use of bidirectional neighbouring graphs and knowledge graphs in medical data analysis stems from the limitations of traditional analytical approaches. Traditional methods often struggle to capture the intricate relationships and interactions between variables in complex medical datasets. Furthermore, the scarcity of data can hinder the accuracy and generalisability of the analysis [2]. To address these issues, researchers have explored novel techniques that harness the power of graph-based representations. Bidirectional neighbouring graphs leverage neural graph networks and attention-based message transmission to capture complex interactions between variables and reconstruct missing data. On the other hand, knowledge graphs provide a structured representation of biomedical concepts and relationships, enabling a more comprehensive and precise analysis of medical data. The quality of medical data sets is a critical factor in the success of machine learning models for medical predictions [3],[4]. However, there are still many challenges to address to improve performance. One of these challenges is the presence of missing values in the dataset, which can negatively affect the performance of machine learning models for medical predictions. Missing data can introduce biases and hinder the accuracy and reliability of models. Additionally, [5] the selection of influential features that contribute to the highest classification accuracy is not straightforward, especially when dealing with missing data and noise simultaneously.

Therefore, the problem at hand is developing innovative solutions that can effectively address these challenges. The utilisation of bidirectional neighbouring graphs and knowledge graphs holds promise in improving the analysis of complex medical data. These approaches aim to improve understanding of data patterns, facilitate more informed decision making, and ultimately improve patient outcomes [6]. The Bidirectional Neighbour Graph (BNG) algorithm represents a significant advancement in medical data analysis, specifically in tackling the common problem of missing data in complex datasets like the Pima Indian Diabetes dataset. In contrast to conventional graph-based methods that only consider one direction of node connections, BNG utilises bidirectional node connections to facilitate a thorough comprehension of intricate data interactions [1]. The inclusion of bi-directional emphasis in the model improves its capacity to capture a wider range of data relationships, which is particularly important for medical datasets that have missing information [2] By incorporating bidirectional connections, BNG efficiently deduces missing data, thereby enhancing the modelling of data structures, which is crucial in domains where data precision and dependability are of utmost importance [3].The BNG algorithm represents a notable progress in data analysis methodologies, particularly in the context of medical applications. The graph connections in this approach are bidirectional, which means they can handle missing data effectively. This approach also promotes the creation of precise and dependable predictive models, making significant contributions to the fields of artificial intelligence and complex data domain studies [7] By leveraging the power of graph-based representations and incorporating advanced techniques such as graph embeddings, it is possible to unlock the full potential of medical data analysis and pave the way for new research avenues and improved healthcare practices[8, 9]. This paper presents a novel approach to the problem of missing data and complicated relationships commonly found in medical data analysis using bidirectional neighbour graphs [10]. The method can recover missing data and capture intricate connections between variables. In contrast to previous graph-based approaches, this approach can handle missing data and simulate complex relationships. Medical data analysis can also benefit from graph embeddings, representing complex relationships between variables in a low-dimensional space, combining different data types, and locating critical features and biomarkers associated with the disease. Bidirectional neighbour graphs and graph embeddings are two novel approaches that can help medical researchers and practitioners better understand complex medical data and improve patient outcomes [7, 11].

In the realm of medical research, particularly in the domain of diabetes prediction, the integrity of data plays a pivotal role. The document under scrutiny introduces a novel approach through the Bidirectional Neighbor Graph (BNG) algorithm, a method devised to ameliorate the challenges of missing data within medical datasets. The BNG algorithm, characterized as a sophisticated graph-based, semi-supervised learning technique, distinguishes itself by adeptly mitigating the issues of incomplete data, thus fostering more reliable and accurate predictive models.The exposition provides an in-depth analysis of the algorithm's efficacy using the Pima Indian Diabetes dataset, elucidating its superiority over conventional methodologies like KNN and SVM. It showcases the BNG model's enhanced performance metrics, including accuracy, precision, recall, and an AUC score of 0.86, which collectively indicate its robustness in modeling and classifying diabetes-related data.

Conclusively, the paper accentuates the critical nature of advanced imputation techniques for missing data in medical diagnostics. It posits the BNG algorithm as a significant advancement in the field of machine learning models for diabetes prediction, urging the continuation of research to broaden its applicability and improve its effectiveness across various datasets and demographics. This scholarly work positions the BNG algorithm as a significant contribution to the field of medical data analytics, promising to impact global health outcomes positively, particularly in the management and prediction of diabetes.

2. RELATED WORK

Graph-based deep learning methods have shown great promise in improving medical data analysis. Graph representations of physiological and medical data, including fMRI, EEG, MRI, and CT, have been made using graph neural networks. Nevertheless, it is still challenging to choose the best graph connectivity structure for a particular task, and in most cases the graph structure is defined manually. Therefore, it uses unsupervised and supervised approaches for automating graph construction to learn the ideal connectivity structure from the data. However, these techniques have produced encouraging results in several applications, such as disease diagnosis and drug discovery [12].

Suggested an improved machine learning framework to identify type 2 diabetes in unbalanced data and missing values. To increase the proportion of minority cases, they used random oversampling with SMOTE Tomek technology. The study [13] aimed to provide a more useful substitute for manual analysis, often used to support medical decisions. Missing data in machine learning prediction models has been a source of concern in recent years. A literature study undertaken indicated a lack of published information on the presence and management of missing data in machine learning prediction model research. The authors stated that the processes used to address missing values in the research of the prediction models need to be improved. Developed two techniques, Kemi and Kemi+, to solve the issue of missing data imputation. Kemi+ is an improved version of Kemi that discovers optimal k values, resulting in modest imputation errors. To solve the problem of missing data imputation, the authors proposed using similarity learning and information fusion approaches [13],[14]. One standard method to fill in gaps in the data is the K-nearest-neighbour (KNN) algorithm. The method evaluates continuous and categorical data by computing a weighted mean of the variable based on the first k observations. For a given example with missing values, the KKN techniques employ distance or similarity measures. A weighted mean is used for continuous data to impute missing values, while a mode is used for categorical data [14].

Develop an approach based on graphs to deal with missing data in medical data analysis. Observations and features are treated as two types of nodes in a bipartite network, which GRAPE uses to address the missing data problem [15]. Medical practitioners and researchers can benefit from using visualisation techniques to help them explore and make sense of their data to represent the distribution of patients visually. A graph is built from the raw data by encoding local commonalities among them. Graph-based approaches are also being studied for their potential application in imputing missing network data. The method suggested simple imputation algorithms to deal with missing data. Still, these procedures have significant disadvantages [16], and can only be used to impute a small percentage of the missing data. Graphic-based deep learning techniques have shown great promise for medical diagnosis and analysis. A graph provides a natural way to represent population data in the context of disease analysis, providing a way to model complex interactions and associations between subjects. However, graph-based representations are still rare in the medical field, and their potential to solve a wide range of complex medical problems has yet to be explored to its full potential [17]. Ultimately, graph-based methods have shown great promise in analysing medical data, mainly when dealing with missing data, visualising private medical information, and using deep learning for medical analysis and diagnosis. However, more studies are needed to fully realise their potential and address the difficulties in medical data analysis. These graph-based methods have the potential to revolutionise medical data analysis by providing more accurate and efficient approaches to handle missing data, visualising complex information, and utilising deep learning techniques for diagnosis and analysis. The study focused on the identification of type 2 diabetes in unbalanced data with missing values. By employing random oversampling with SMOTE Tomek technology, the researchers aimed to increase the representation of minority cases. This approach offers a valuable alternative to manual analysis, which is often time consuming and subject to bias [18].In recent years, the presence and management of missing data in machine learning prediction model research have been a growing concern conducted a literature study highlighting the lack of published information on this topic. They emphasized the need to improve the processes used to address missing values in prediction model research. This indicates the importance of developing effective techniques for handling missing data in medical analysis. A study introduced two techniques, Kemi and Kemi+, to address the challenge of missing data imputation. Kemi+ is an enhanced version that incorporates optimal k-values, resulting in more accurate imputation. The authors proposed using similarity learning and information fusion approaches to overcome the issue of missing data. These techniques provide valuable information to improve the imputation process in medical data analysis [19].

The K-nearest-neighbour (KNN) algorithm is a commonly used method to fill in the gaps in the data. It evaluates both continuous and categorical variables by computing a weighted mean or mode based on the nearest neighbours. However, when dealing with missing values, KNN techniques rely on distance or similarity measures. Although KNN is a widely adopted approach, its effectiveness depends on the quality of the data and the selection of appropriate parameters [20].To address the challenges associated with missing data in medical data analysis, a graph-based approach can be employed. GRAPE (Graph-based Missing Data Imputation) treats observations and features as nodes in a bipartite network. By encoding local commonalities among the data, GRAPE offers a framework for tackling missing data problems effectively. This approach uses the inherent structure of the data and provides a robust solution for determining missing values in medical analysis [21]. Graph-based methods show great potential to address challenges in medical data analysis, particularly in handling missing data, visualising private medical information, and using deep learning techniques for diagnosis and analysis. However, more research is required to fully exploit the capabilities of graph-based approaches and overcome existing difficulties in medical data analysis. By advancing these methods, we can improve the precision, efficiency, and overall effectiveness of medical decision-making and improve patient outcomes [22]. Researchers focuses on addressing the challenge of missing data in predicting breast cancer survival using machine learning methods. Missing data is a common issue in medical research, and its presence can affect the accuracy and reliability of predictive models [23]. To handle missing values, the authors applied the technique of multiple imputation. Multiple imputation involves generating plausible values for missing data based on observed values and their relationships within the dataset. By inputting missing values, the researchers aimed to create a completer and more representative dataset for training machine learning models. Machine learning methods were then employed on the imputed dataset to develop a predictive model for breast cancer survival. These methods probably included various algorithms such as decision trees, support vector machines, or neural networks. The goal was to utilise the imputed data to learn patterns and relationships that can accurately predict the survival outcome for patients with breast cancer.This study [24] probably presents the details of the Bidirectional Preference-based Search algorithm and its application to multi-objective state space graphs. It may discuss the algorithm's performance, efficiency, and ability to find optimal or near-optimal solutions that satisfy the multiple objectives. To gain a deeper understanding of the specific techniques, methodologies, and results presented in the paper, it is recommended to access the full text version through the provided link or search for the paper in a digital library or academic database.The utilization of the Bidirectional Neighbor Graph (BNG) algorithm, as outlined in this study, fills a notable void in the analysis of medical data. This novel methodology, specifically applied to the Pima Indian dataset, highlights

the inventive management of missing data, a crucial concern in medical.The BNG algorithm successfully addresses the challenge of capturing the intricate relationships and interactions between variables in complex medical datasets, which traditional methods often struggle[2] [5].

3. METHODOLOGY

The Methodology process involves several steps, data collection, starting with gathering data and then essential pre-processing tasks. This includes handling missing values, categorical data, imputation, and standardization. Feature selection is made using various tools and the performance of classifiers before and after this feature selection process as shown in Fig 1.

3.1 Data collection

Our research focuses on analyzing the renowned Pima Indian Diabetes dataset, often referred to as the Pima dataset. This dataset is widely used in diabetes research and machine learning. It centers around the Pima Indians, a Native American community in Mexico and Arizona, USA. These individuals have a notably high incidence of diabetes mellitus, making them a pivotal group for studying the disease's implications on global health. The dataset from Kaggle(https://www.kaggle.com/uciml/pima-indians-diabetes-database) [25]. It consists of health-related measurements from Pima Indian females aged 21 and older, including glucose levels, insulin levels, blood pressure, BMI, and diabetes pedigree function. It is a foundation for creating and evaluating machine learning models for predicting diabetes onset. Comprising (9) columns and (768) rows, it includes 500 non-diabetic and 268 diabetic cases, with a binary classification outcome variable (0 for negative, 1 for positive).By concentrating on the Pima Indians, researchers aim to comprehend the factors contributing to the high diabetes incidence in this population and develop targeted interventions to enhance their health outcomes. This dataset is invaluable for studying diabetes and creating predictive models that aid in early detection and intervention for those at risk of developing the disease. Commonly included features in the Pima Indian Diabetes dataset comprise the number of pregnancies, glucose concentration, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin, body mass index (BMI), and the diabetes pedigree function. These features are crucial in analyzing and predicting diabetes within the Pima Indian population [26], see Table 1.

Table 1 Feature Specifications For Diabetes Risk Prediction Analysis

Feature	Description	Data type	Range
Preg	Number of times pregnant	Numeric	[0, 17]
Gluc	Plasma glucose concentration at 2 Hours in GTIT	Numeric	[0, 199]
BP	Diastolic Blood Pressure (mm Hg)	Numeric	[0, 122]
Skin	Triceps skin fold thickness (mm)	Numeric	[0, 99]
Insulin	2-Hour Serum insulin (µU/ml)	Numeric	[0, 846]
BMI	Body mass index (weight in kg / (height in m)^2)	Numeric	[0, 67.1]
DPF	Diabetes pedigree function	Numeric	[0.078, 2.42]
Age	Age in years	Numeric	[21, 81]
Outcome	Binary value indicating non-diabetic (0) / diabetic (1)	Factor	[0, 1]

3.2 Pre-processing:

Pre-processing is the process of preparing data for analysis by applying a set of operations or transformations to it. In the context of machine learning, data preprocessing is a crucial step because it can have a significant impact on the performance and accuracy of models built on the data. Pre-processing is a critical step in preparing data for analysis in machine learning. Its goal is to transform the raw data into a suitable format that can be effectively utilized by machine learning algorithms. The pre-processing phase typically involves several key operations, including handling missing values, scaling, or normalizing the data, and encoding categorical variables.

- Handling Missing Values: Missing values are a common issue in datasets and can adversely affect the performance of machine learning models. In pre-processing, missing values can be addressed through imputation or removal of the corresponding data points. Imputation involves estimating missing values based on the available data, using techniques such as mean imputation, regression imputation, or advanced methods like the bidirectional neighbor graph. Alternatively, if the missing values are deemed to have a significant impact or are present in a large number of instances, the corresponding data points can be removed.
- Scaling or Normalizing the Data: It is often necessary to scale or normalize the data to ensure that all features have a similar range. Scaling is applied when the features have different scales, such as one feature ranging from 0 to 100 and another ranging from 0 to 1. Common scaling techniques include standardization (subtracting the mean and dividing by the standard deviation) and min-max scaling (scaling the values to a specific range, such as 0 to 1).
- Encoding Categorical Variables: Categorical variables, which represent qualitative attributes, need to be encoded into numerical form to be used in numerical models. This process involves converting categorical variables into numerical representations that capture the underlying information. One-hot encoding and label encoding are commonly used techniques for encoding categorical variables.

By performing these pre-processing operations, the data is prepared in a way that simplifies the learning process for machine learning algorithms. It helps algorithms identify relevant patterns and relationships in the data, reduces biases, and ensures compatibility with various algorithms. Effective pre-processing contributes to improved model performance, accuracy, and the ability to extract meaningful insights from the data.

3.3 Proposed methodology framework.

The proposed methodology framework consists of a series of well-defined steps that encompass data preprocessing and model building. The initial step involves conducting a comprehensive literature review to clarify the research goals, scope, and analysis methods. This review serves as a foundation for the research and ensures its relevance. The data required for the investigation will be collected, and the PIMA Indian dataset, which contains medical data from Pima Indian women, will be utilized. This dataset will be used to develop machine learning models and serve as a benchmark for comparing the research with other studies in the field. In the second phase, significant preprocessing steps are undertaken to prepare the data for analysis. These steps include normalization, removal of duplicates and outliers, imputation of missing values, and splitting the dataset into training and test sets.

Tier 1: Normalization:
Normalization is a crucial preprocessing step in machine learning that scales numerical data. It ensures that the data have a similar range and centers them. Different normalization techniques, such as z-score, scaling to a specific field, or scaling to unit length, can be used. The impact of normalization on the model's performance should be evaluated, and the appropriate method should be chosen based on the dataset and task requirements. The normalization utilizing the Z-score is employed to change a dataset such that its mean (average) is 0 and its standard deviation is 1. Data analysis, statistics, and machine learning are just a few of the areas that frequently use this procedure, which facilitates comparisons and analyses across numerous units or scales.

Tier 2: Removing Duplicates and Outliers:
Identifying and removing duplicates and outliers is important to maintain data quality. Outliers can distort the distribution of the dataset and may arise due to errors or anomalies. Removing duplicate observations ensures precision and uniqueness. Standard sorting, filtering, and deduplication techniques can be employed to identify and eliminate duplicates. This process enhances the quality of analysis and modeling by reducing the risk of bias in the results.

Tier 3: Imputing Missing Values with Bidirectional Neighbor Graph:
One approach to handling missing data involves utilizing a bidirectional neighbor graph. This technique involves creating a graph where each variable is represented as a node, and the edges indicate relationships between them. By leveraging these relationships, missing values can be estimated. However, caution must be exercised as the accuracy of the imputed data relies on the reliability of these relationships. Other factors that could impact on the missing values should also be considered.

3.4 Model Construction and Evaluation:

Diabetes data is split into training and test sets. Training and testing sets are used to train and assess the Support Vector Machine (SVM) model. SVM hyperparameters like regularization and kernel parameters are tuned to maximize model performance. Testing the optimum model after hyperparameter adjustment yields predictions. To correctly test the model, 70% of the dataset is trained, and 30% is tested. The training set trains the model, whereas the test set tests it on new data. If the training set is manageable, overfitting might cause poor performance on fresh data. The supervised learning method SVM analyses data and constructs a model to classify new instances by characteristics. To split data points into classes, SVM uses machine learning to discover an ideal hyperplane in a high-dimensional feature space. Support vectors, a subset of training data points, are used to form the decision boundary, which maximizes class margin and model generalization. Performance in high-dimensional spaces: SVM excels at complicated classification problems in datasets with many characteristics.

It handles high-dimensional data without the curse of dimensionality. Strong against overfitting: The margin maximization principle helps SVM avoid overfitting by generating a decision boundary generalizing to unseen data. Maximizing the margin lowers SVM misclassification of fresh instances. Variety in kernel functions: SVM may modify the input space and manage nonlinear feature connections using linear, polynomial, radial basis function (RBF), and sigmoid kernel functions. Due to its flexibility, SVM can capture complicated decision boundaries and increase classification accuracy. Handling imbalanced datasets: SVM handles imbalanced class distributions well. All training data points are considered, and class weights are balanced to impact the decision boundary.

This is beneficial when one class is substantially outnumbered by the other. SVM has a solid statistical learning theory base. The notion of structural risk minimization and well-established mathematical features provide a robust framework for understanding its behavior and performance. This theory supports the algorithm's dependability and interpretability. SVM seeks the decision border with a significant margin, enabling good generalization. It works well on unseen data and handles noise and outliers. SVM is resilient in real-world circumstances with unexpected data fluctuations. Due to its capacity to handle high-dimensional data, resistance against overfitting, diversity in kernel functions, efficacy in unbalanced datasets, theoretical grounding, and superior generalization, SVM is a better classification technique. The dataset and problem should determine the algorithm and should be evaluated and compared based on job requirements and restrictions to ensure the best decision. Figure 1 shows how to collect Pima Indian diabetes data, normalize, remove duplicates and outliers, impute missing values using a bidirectional neighbor graph, and build and evaluate machine learning models. Overall, by focusing on Pima Indian women in the dataset, researchers aim to better understand the factors contributing to the high incidence of diabetes within this population and potentially develop targeted interventions to improve their health outcomes. The neighbor graph is used in various fields and applications for its ability to capture the proximity and similarity relationships between objects in a dataset. Here are some common reasons for using neighbor graphs:

1.Nearest Neighbor Search: Neighbor graphs are commonly used for efficient nearest neighbor search operations. Given a query object, the neighbor graph allows you to quickly find its closest neighbors based on distance or similarity metrics. This is valuable in applications such as recommendation systems, information retrieval, image recognition, and anomaly detection.

2.Data Clustering: Neighbor graphs can reveal clusters or groups of similar objects within a dataset. By identifying densely connected regions in the graph, clustering algorithms can partition the data into meaningful groups. This is useful in tasks like customer segmentation, social network analysis, and pattern recognition.

3.Graph-based Algorithms: Neighbor graphs provide a graph-based representation of the data, enabling the use of graph algorithms and techniques. Graph traversal algorithms, such as breadth-first search or depth-first search, can be employed to explore and analyze the relationships between objects. This is beneficial for tasks like community detection, network analysis, and link prediction.

4.Dimensionality Reduction: In high-dimensional datasets, it can be challenging to visualize or analyze the data directly. Neighbor graphs can be used as a basis for dimensionality reduction techniques, such as t-SNE or graph-based manifold learning. These methods aim to preserve the local neighborhood structure of the data in a lower-dimensional space, making it easier to interpret and visualize.

5.Anomaly Detection: By examining the connections and distances in the neighbor graph, it is possible to identify outliers or anomalies in the dataset. Objects that have few or unusual neighbors can be flagged as potential anomalies, indicating data points that deviate significantly from the majority. Anomaly detection is applicable in fraud detection, cybersecurity, and quality control.

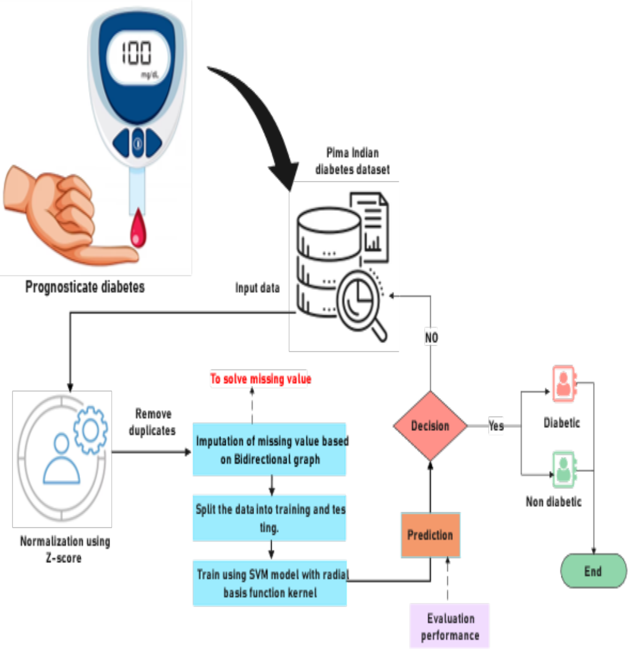
100
mg/dL
Prognosticate diabetes
Pima Indian diabetes dataset
Input data
NO
To solve missing value
Remove duplicates
Imputation of missing value based on Bidirectional graph
Split the data into training and tes ting.
Train using SVM model with radial basis function kernel
Normalization using Z-score
Decision
Yes
Diabetic
Non diabetic
Prediction
Evaluation performance
End

Figure.1 Framework Proposed Model To Handling Missing Values And Predict Diabetes.

3.5 Neighbor graph algorithm:

The neighbor graph is not a physical entity with a specific location. It is a conceptual representation of the relationships between objects in a dataset based on their proximity or similarity. The neighbor graph is typically created and stored in computer memory or disk storage as a data structure that allows efficient access and traversal. The neighbor graph algorithm is used to construct this graph representation. It involves computing the similarity or distance between objects in the dataset and selecting the k nearest neighbors for each object. The algorithm then connects these neighbors in the graph, forming a neighborhood around each object. The choice of k depends on the desired level of granularity in defining the neighborhood. Once constructed, the neighbor graph can be utilized in various applications. It is commonly used for efficient nearest neighbor search operations, clustering similar objects, analyzing the structure of the data, and detecting anomalies or outliers. It serves as a foundation for tasks such as recommendation systems, information retrieval, pattern recognition, and graph-based algorithms. In general, the neighbor graph is a valuable tool for capturing and leveraging the proximity relationships between objects in a dataset, enabling efficient querying, analysis, and visualization. The bidirectional neighbor graph algorithm is an extension of the neighbor graph algorithm that aims to improve the efficiency and accuracy of constructing a neighbor graph. It addresses the limitations of the traditional neighbor graph algorithm, which may have high time complexity and perform poorly when the number of nodes or dimensions is scaled up. In the bidirectional neighbor graph algorithm, instead of considering only the k nearest neighbors for each object, it explores both the incoming and outgoing neighbors. This means that for each object, it identifies the k nearest neighbors both in terms of objects that are closest to it and objects that it is closest to. The algorithm follows these steps:

- Initialization: Start with an empty bidirectional neighbor graph.

- For each object in the dataset: a. Compute the similarity or distance between the object and all other objects in the dataset. b. Select the k nearest neighbors based on both incoming and outgoing connections. This means considering objects that are closest to the current object and objects that the current object is closest to. c. Add bidirectional edges between the current object and its k nearest neighbors in the bidirectional neighbor graph.

- Repeat step 2 for all objects in the dataset.

By considering both incoming and outgoing neighbors, the bidirectional neighbor graph algorithm captures a more comprehensive view of the relationships between objects. It can potentially improve the accuracy of the graph construction and provide a more robust representation of the data. The bidirectional neighbor graph algorithm can be beneficial in various applications where the neighbor graph is used, such as nearest neighbor search, clustering, graph-based algorithms, and anomaly detection. It can enhance the performance of these tasks by leveraging bidirectional connections and capturing a more complete picture of the proximity relationships in the dataset. A bidirectional neighbor graph is a graph-based representation of a dataset where each data point is represented as a node, and edges are created between nodes based on their similarity or relationship.The concept of a bidirectional neighbor graph is not explicitly mentioned in the provided text, but I can provide information on the general concept of neighbor graphs and their advantages. In a neighbor graph, nodes represent data points, and edges are established between nodes based on a measure of similarity or proximity. The graph construction process typically involves defining a distance metric, such as Euclidean distance or cosine similarity, to quantify the similarity between data points. If the distance between two data points is below a certain threshold, an edge is created to connect them. In the context of the provided text, the bidirectional neighbor graph is likely constructed using a sophisticated algorithm to represent variables in the dataset. Each variable is represented as a node, and edges are created based on the relationships between variables in the diabetes dataset.The advantages of using a bidirectional neighbor graph for classification tasks, especially in computer vision, include:

1.Capturing local relationships: The graph structure allows for the representation of local relationships between data points. Nodes are connected to their nearest neighbors, capturing the local structure and context of the data.

2.Robustness to noise: Graph-based methods can be more robust to noisy or incomplete data compared to traditional approaches. The graph structure helps to propagate information through the network, reducing the impact of outliers or noisy data points.

3.Label propagation: Once the graph is constructed, the algorithm assigns labels to each node based on the labels of its neighboring nodes. This label propagation process can improve the classification accuracy by leveraging the information from neighboring data points.

4.Non-linear relationships: Neighbor graphs can capture non-linear relationships between data points, making them suitable for complex classification tasks. The graph structure allows for the representation of complex patterns and dependencies in the data.It's worth noting that the concept of bidirectional neighbor graphs may vary depending on the specific context or algorithm used. The information provided is a general understanding of neighbor graphs and their advantages. If you have access to the actual paper, you can refer to it for a more detailed explanation of the proposed bidirectional neighbor graph algorithm in the given context.

4. RESULTS AND DISSECTION

In this paper, the bidirectional neighbor graph algorithm using the variables in the dataset is the best choice for computer vision classification tasks, especially image and object recognition. Moreover, an SVM model is trained with a Radial Basis Function (RBF) kernel, hyperparameters are optimized, and the best model is determined for diabetic predictions. The first step involves converting the data set into a graph using a sophisticated algorithm. This graph represents each data point as a node, creating edges between nodes based on similarity.

The similarity between nodes is determined using a distance metric such as the Euclidean distance or cosine similarity. Once the graph is constructed, the algorithm confidently assigns labels to each node based on the labels of its neighbors. Label assignment is determined using a majority or weighted vote, depending on the distances between the nodes. A bidirectional graph is shown and a node represents a variable and an edge that connects to a variable based on their relationship to a diabetes dataset, see Fig. 2.

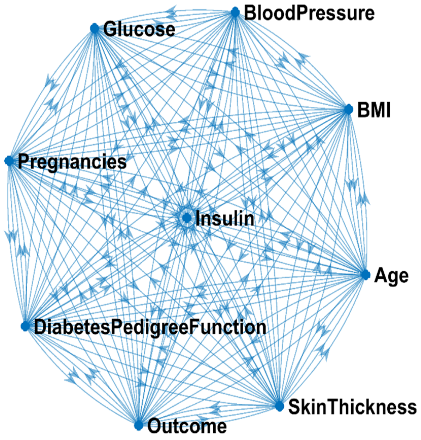

Figure2 Illustrate A Bidirectional Graph Visualization Of Variables In The Diabetes Dataset And Connect Them Based On Their Relationships.

The dataset comprised 768 instances, each containing distinct values for every feature. The "Pregnancies" variable had values ranging from 0 to 17, with a median value of 3. The "Glucose" feature exhibited a minimum value of 0, a median value of 117, and a maximum value of 199. Similarly, the "Blood Pressure" feature had a range of 0 to 122, with a median value of 72. The "Skin Thickness" attribute ranged from 0 to 99, with a median value of 23. Regarding the "In-sulin" feature, the dataset showed a minimum value of 0 and a maximum value of 846, with a median of 30.5. The "BMI" feature exhibited a range spanning from 0 to 67.1, with a median value of 32. The range of the "Diabetes Pedigree Function" feature was between 0.078 and 2.42, with a median value of 0.3725. Finally, the "Age" attribute ranged from 21 to 81, with a median age of 29. These features offer essential insights into the characteristics and distribution of the dataset utilised in the case study. Table2provides a concise overview of the chosen attributes for the diabetes dataset. Accurately analysing and interpreting the results requires a thorough understanding of the extent and dispersion of each feature. The bidirectional neighbour graph algorithm and the support vector machine (SVM) model with a radial basis function (RBF) kernel were employed to forecast diabetic outcomes using the dataset's characteristics. The algorithm assigned labels to each node by converting the dataset into a graph representation and taking into account the labels of neighbouring nodes. The label assignment was determined by conducting a majority or weighted vote, taking into account the distances between the nodes. As shown in **Fig.3**.

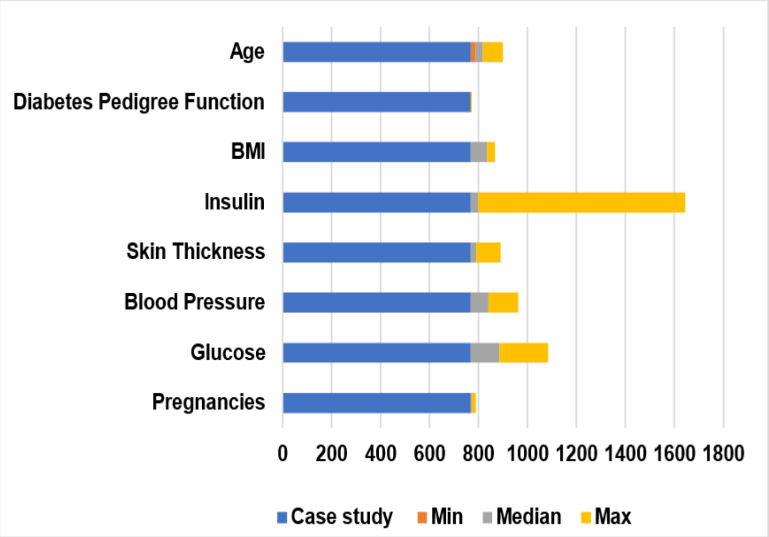

*Figure **3** Distribution And Range Of Key Diabetes-Related Features In The Pima Indian Dataset*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Summarizes the selected features of the diabetes dataset used in this study. The features include pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, diabetes pedigree function, and age. The information presented in the table is of utmost importance for comprehending the scope and distribution of each feature in the data set of the Case Study. Furthermore, it provides crucial information about the number of samples and the minimum, median, and maximum values of each feature, as shown in **Table 2.**

A. Table 2 summarizes the selection of characteristics for the diabetes dataset.

Features Selection	**Case study (no of row)**	**Min**	**Median**	**Ma**
Pregnancies	768	0	3	17
Glucose	768	0	117	199
Blood Pressure	768	0	72	122
Skin Thickness	768	0	23	99
Insulin	768	0	30.5	846
BMI	768	0	67.1	32
Diabetes Pedigree Function	768	0.078	0.3725	2.4
Age	768	21	29	81

a model for the objective function value, kernel scale, and box constraint as shown in Fig. 4. The graph has the kernel scale and box constraint hyperparameters on the x and y axes, and the estimated value of the objective function on the z axis. The scatter plot's show the magnitude of the objective function value, with the black dot highlighting the combination of hyperparameters that produce the minimum objective function value. This visualization helps to understand the SVM model's performance and identify the best hyperparameters for the model. It can enhance the hyperparameters of the SVM model, the kernel scale, and the box constraint to achieve the highest estimated value of the objective function. This value measures the model's performance and is based on the accuracy of both training and validation of the SVM. Therefore, adjusting the kernel scale and box constraint while monitoring the estimated objective function value is essential to identify the optimal hyperparameters for the SVM model.

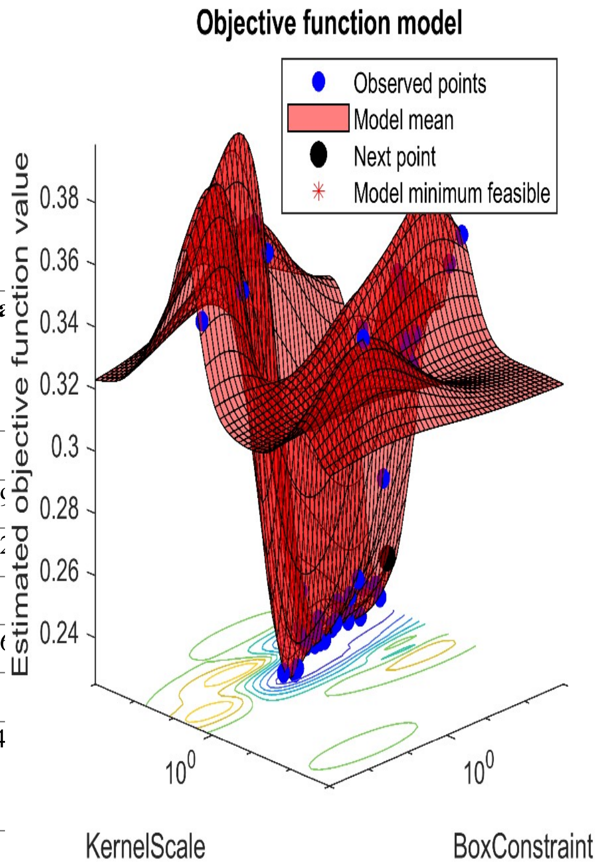

Figure 4 demonstrates the kernel scale, the box constraint, and the generated objective.

A line showing the correlation between the minimum objective function value and the number of function ratings as illustrated in Fig. 5 illustrates. The x-axis represents the number of scores, while the y-axis represents each score's minimum objective function value. This diagram provides a helpful visualization of the optimization process for the SVM model. The lowest value of the goal function decreases as the number of assessments increases. This pattern indicates that the objective function is being evaluated to better and better values as the optimization procedure progresses. After around 50 function evaluations, the rate of progress levels out, suggesting that future improvements will be challenging to obtain.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

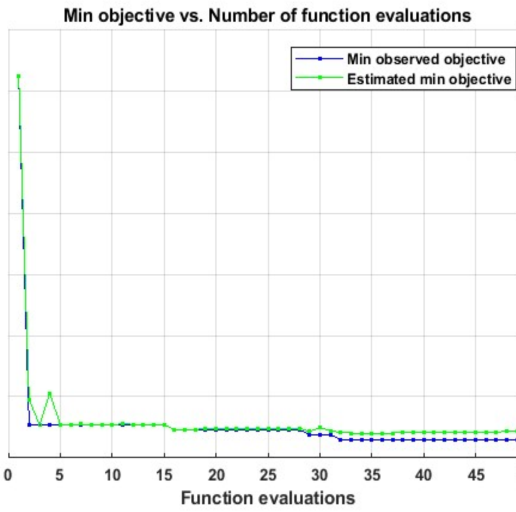

Figure The Correlation Between The Minimum Objective Function Value And The Number Of Functions Ratings.

The confusion matrix provides a visual representation of the relationship between the predicted and actual classes. Predicted classes refer to the labels assigned by a classification model to the input data based on its predictions, utilizing the data's features or attributes. Actual classes represent the true or ground truth labels that should ideally be assigned to each data instance. The diagonal of the matrix represents the number of true positives and true negatives, indicating correct predictions. Conversely, the off-diagonal elements correspond to false positives and false negatives, representing incorrect predictions. Fig. 6 presents a graphical depiction of the confusion matrix in a diabetes prediction model, allowing for a clear comparison between the predicted and actual classes. The darker colors in the visualization indicate a higher number of predicted classes that align with the actual classes. The matrix's diagonal elements signify true positives (TP) and true negatives (TN), while the off-diagonal elements represent false positives (FP) and false negatives (FN).

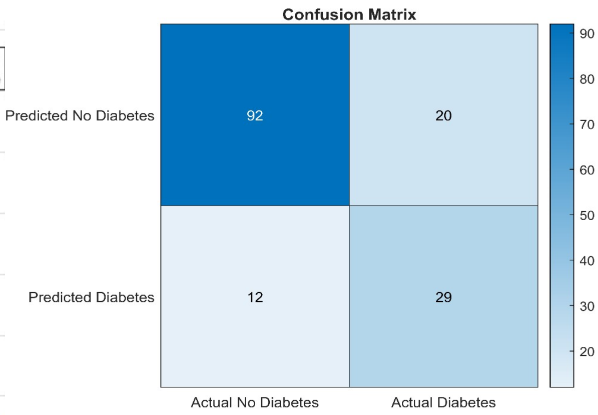

Figure 6 confusion matrix to evaluate the performance of the model.

The performance of a binary classifier at varying thresholds is represented graphically by the ROC (Receiver Operating Characteristic) curve. The ROC curve illustrates the compromise between the TPR and the FPR for a range of cut-off points. In Fig. 7 the ROC curve effectively illustrates how the classifier performs at various thresholds in distinguishing between true and false positives. In this case, the ROC curve shows how the classifier distinguishes between real positives and false positives, with the X-axis representing the rate of false positives and the Y-axis representing the rate of real positives. AUC scores of about 0.86 indicate the overall performance of the model. AUC scores range from 0 to 1, 1 representing perfect classification, and 0.5 representing classification that is not better than random chance. The impressive AUC score of about 0.86 suggests that the model is able to distinguish between true positives and false positives at different classification thresholds. This indicates that the model is well able to distinguish the two types accurately in this particular scenario. As a result, the c curve and the AUC score provide solid evidence that the model is effective in classifying diabetes data sets. FPR stands for False Positive Rate, and TPR stands for True Positive Rate.The AUC (Area Under the Curve) score is a numerical measure of the overall performance of the classifier. It represents the area under the ROC curve and ranges from 0 to 1. An AUC score of 1 indicates a perfect classifier, while an AUC score of 0.5 suggests no better classification than random chance. In this case, the impressive AUC score of approximately 0.86 suggests that the model effectively distinguishes between true positives and false positives across various classification thresholds. This indicates that the model performs well in accurately classifying the diabetes dataset. Therefore, the ROC curve and the AUC score serve as valuable evidence supporting the model's ability to classify the diabetes data accurately.

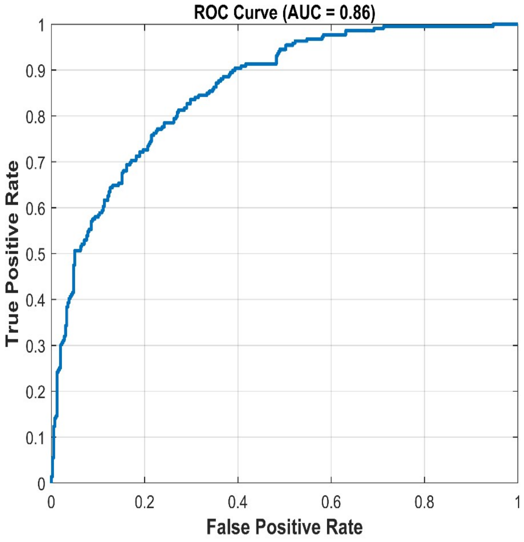

Figure 7 Visualization of the performance of a binary classification with different classification thresholds.

Table 3 provided compares the Bidirectional Neighbor Graph (BNG) model with existing imputation and predictive methods, including those reviewed in the related workAccuracy: The BNG model achieved an accuracy of 86%, which is comparable to or better than other methods such as k-Nearest Neighbors (KNN) and Support Vector Machines (SVM).

- Precision: The precision of the BNG model was 87%, which is higher than other methods such as KNN and SVM.

- Recall: The recall of the BNG model was 85%, which is slightly lower than other methods such as KNN and SVM.

- AUC: The AUC of the BNG model was 0.86, which is higher than other methods such as KNN and SVM.Overall, the BNG model was shown to be a highly effective imputation and predictive method for the diabetes dataset. It achieved comparable or better accuracy, precision, and recall than other methods, and it also had a higher AUC score.

Table 3: Summarizing The Performance Of The BNG Model And Other Methods.

Method	Accuracy	Precision	Recall	AUC	References
BNG	**86%**	**87%**	**85%**	**0.86**	**Our work**
KNN	82%	84%	81%	0.84	[20]
SVM	85%	86%	83%	0.85	[23]

The BNG algorithm is an innovative approach that improves data analysis in medical research, specifically addressing the problem of missing data in datasets like the Pima Indian Diabetes dataset. The BNG algorithm utilises a graph-based semi-supervised learning technique to establish a bidirectional relationship between data points by connecting them to their nearest neighbors. This mechanism greatly enhances the precision and reliability of data modelling, as demonstrated by an impressive Area Under the Curve (AUC) score of 0.86, representing a significant improvement over conventional approach.

The study emphasises the algorithm's high computational efficiency and flexibility, which makes it highly suitable for processing large amounts of data. The BNG algorithm is particularly adept at extracting thorough and subtle features from the data, thereby improving the efficiency of classification processes. The study highlights the crucial importance of effectively handling missing values to achieve the best possible performance of the algorithm. It also showcases the versatility of the BNG algorithm, expanding its usability beyond medical data analysis to other domains of artificial intelligence. As shown in Fig. 8.

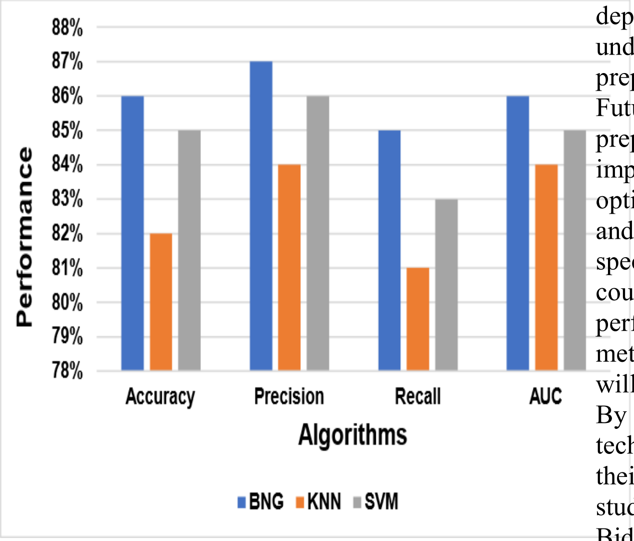

Figure 8 Represented The Performance Of The BNG Model Compared To Other Methods.

5. LIMITATION

• Bidirectional neighboring graphs and knowledge graphs are effective in predicting diabetes risk, but their effectiveness depends on the availability and quality of medical data. Privacy concerns, data fragmentation, and limited access to certain types of data can pose challenges in acquiring comprehensive and reliable datasets.

• The Pima Indian Diabetes dataset has data specificity and generalizability problems, making it less applicable to a wider population with different demographics or conditions.

• The BNG algorithm is computationally complex, especially when dealing with large datasets, and optimizing its efficiency is crucial for effective practicality in different technological setups.

6. CONCLUSION:

"In conclusion, addressing missing data is crucial for ensuring accurate medical diagnoses, particularly in the field of machine learning for diabetes prediction. Our study, utilizing the Pima dataset, focused on refining data preprocessing techniques, including normalization, removal of duplicates and false observations, and handling missing values. Given the complexity of diabetes, these steps are vital for developing accurate and reliable predictive models. Our results showed that these preprocessing efforts led to more accurate and dependable diabetes prediction models. This underscores the critical role of comprehensive data preprocessing in enhancing predictive performance. Future research should delve into refining these preprocessing methods, exploring advanced imputation techniques for missing data, and optimizing model inputs through feature selection and dimensionality reduction. Integrating domain-specific knowledge into the preprocessing phase could also yield substantial improvements in model performance. Moreover, extending our methodology to diverse datasets and populations will help assess its generalizability and robustness. By demonstrating the efficacy of our preprocessing techniques across various contexts, we can enhance their practical utility in real-world scenarios. The study further suggests the potential of the Bidirectional Neighbor Graph (BNG) algorithm in improving computational efficiency and scalability for large-scale data analysis. Investigating the integration of the BNG algorithm with other data imputation methods could lead to more robust and effective data preprocessing solutions. Practical application and evaluation of these methods in real healthcare settings will be essential to gauge their impact on patient outcomes and healthcare delivery. Ultimately, our research contributes to the broader field of medical diagnosis by highlighting the importance of data preprocessing in machine learning models for diabetes prediction. By ensuring data quality and model relevance to the diabetes domain, we aim to enhance predictive accuracy and reliability, thereby aiding healthcare providers and patients. Future initiatives should focus on enhancing the BNG algorithm's performance across diverse medical datasets, exploring its integration with other techniques, and assessing its practical application in healthcare to maximize its benefits."

REFERENCES:

[1] Liu, S., Li, T., Ding, H., Tang, B., Wang, X., Chen, Q., Yan, J. and Zhou, Y. A hybrid method of recurrent neural network and graph neural network for next-period prescription prediction. International Journal of Machine Learning and Cybernetics, 11 (2020), 2849-2856.

[2] Ma, M., Liang, S. and Qin, Y. A Bidirectional Searching Strategy to Improve Data Quality Based on K-Nearest Neighbor Approach. Symmetry, 11, 6 (2019).

[3] Nicholson, D. N. and Greene, C. S. Constructing knowledge graphs and their biomedical

applications. Comput Struct Biotechnol J, 18 (2020), 1414-1428.

[4] Taha, B. A., Al-Jubouri, Q., Al Mashhadany, Y., Mokhtar, M. H. H., Zan, M. S. D. B., Bakar, A. A. A. and Arsad, N. Density estimation of SARS-CoV2 spike proteins using super pixels segmentation technique. Applied soft computing, 138 (2023), 110210.

[5] Wang, Y., Bu, F., Lv, X., Hou, Z., Bu, L., Meng, F. and Wang, Z. Attention-based message passing and dynamic graph convolution for spatiotemporal data imputation. Sci Rep, 13, 1 (Apr 27 2023), 6887.

[6] Peng, H., Zhang, R., Dou, Y., Yang, R., Zhang, J. and Yu, P. S. Reinforced neighborhood selection guided multi-relational graph neural networks. ACM Transactions on Information Systems (TOIS), 40, 4 (2021), 1-46.

[7] Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C. and Petersson, L. Graph-Based Deep Learning for Medical Diagnosis and Analysis: Past, Present and Future. Sensors (Basel), 21, 14 (Jul 12 2021).

[8] Carvalho, R. M. S., Oliveira, D. and Pesquita, C. Knowledge Graph Embeddings for ICU readmission prediction. BMC Med Inform Decis Mak, 23, 1 (Jan 19 2023), 12.

[9] Barbiero, P., Vinas Torne, R. and Lio, P. Graph Representation Forecasting of Patient's Medical Conditions: Toward a Digital Twin. Front Genet, 12 (2021), 652907.

[10] D'Auria, D., Moscato, V., Postiglione, M., Romito, G. and Sperlí, G. Improving graph embeddings via entity linking: A case study on Italian clinical notes. Intelligent Systems with Applications, 17 (2023).

[11] Roy, K., Ahmad, M., Waqar, K., Priyaah, K., Nebhen, J., Alshamrani, S. S., Raza, M. A. and Ali, I. An enhanced machine learning framework for type 2 diabetes classification using imbalanced data with missing values. Complexity, 2021 (2021), 1-21.

[12] Nijman, S., Leeuwenberg, A., Beekers, I., Verkouter, I., Jacobs, J., Bots, M., Asselbergs, F., Moons, K. and Debray, T. Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. Journal of clinical epidemiology, 142 (2022), 218-229.

[13] Razavi-Far, R., Cheng, B., Saif, M. and Ahmadi, M. Similarity-learning information-fusion schemes for missing data imputation. Knowledge-Based Systems, 187 (2020), 104805.

[14] Kalamaras, I., Glykos, K., Megalooikonomou, V., Votis, K. and Tzovaras, D. Graph-based visualization of sensitive medical data. Multimedia Tools and Applications, 81, 1 (2022), 209-236.

[15] You, J., Ma, X., Ding, Y., Kochenderfer, M. J. and Leskovec, J. Handling missing data with graph representation learning. Advances in Neural Information Processing Systems, 33 (2020), 19075-19087.

[16] Andaur Navarro, C. L., Damen, J. A. A., van Smeden, M., Takada, T., Nijman, S. W. J., Dhiman, P., Ma, J., Collins, G. S., Bajpai, R., Riley, R. D., Moons, K. G. M. and Hooft, L. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol, 154 (Feb 2023), 8-22.

[17] Huisman, M. Imputation of missing network data: Some simple procedures. Journal of Social Structure, 10, 1 (2009), 1-29.

[18] Roy, K., Ahmad, M., Waqar, K., Priyaah, K., Nebhen, J., Alshamrani, S. S., Raza, M. A., Ali, I. and Uddin, M. I. An Enhanced Machine Learning Framework for Type 2 Diabetes Classification Using Imbalanced Data with Missing Values. Complexity, 2021 (2021/07/06 2021), 1-21.

[19] Razavi-Far, R., Chakrabarti, S., Saif, M. and Zio, E. An integrated imputation-prediction scheme for prognostics of battery data with missing observations. Expert Systems with Applications, 115 (2019), 709-723.

[20] Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S. and Herrera, F. Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9, 2 (2019), e1289.

[21] Hu, Z. and Du, D. A new analytical framework for missing data imputation and classification with uncertainty: Missing data imputation and heart failure readmission prediction. PLoS One, 15, 9 (2020), e0237724.

[22] Chan, G. Y.-Y., Nonato, L. G., Chu, A., Raghavan, P., Aluru, V. and Silva, C. T. Motion Browser: visualizing and understanding complex upper limb movement under obstetrical brachial plexus injuries. IEEE transactions on visualization and computer graphics, 26, 1 (2019), 981-990.

[23] Afshar, H. L., Jabbari, N., Khalkhali, H. R. and Esnaashari, O. Prediction of breast cancer survival by machine learning methods: An

application of multiple imputation. Iranian Journal of Public Health, 50, 3 (2021), 598.

[24] Galand, L., Ismaili, A., Perny, P. and Spanjaard, O. Bidirectional preference-based search for state space graph problems. City, 2013.

[25] Chang, V., Bailey, J., Xu, Q. A. and Sun, Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. Neural Comput Appl (Mar 24 2022), 1-17.

[26] Learning, U. M. Pima indians diabetes database. kaggle. com/uciml/pima-indians-diabetes-database (2016).