





Article

Fusion of Graph and Tabular Deep Learning Models for Predicting Chronic Kidney Disease

Patike Kiran Rao ^{1,*}, Subarna Chatterjee ¹, K Nagaraju ², Surbhi B. Khan ^{3,*}, Ahlam Almusharraf ⁴
and Abdullah I. Alharbi ⁵

¹ Department of Computer Science and Engineering, Faculty of Engineering, MS Ramaiah University of Applied Sciences, Bengaluru 560054, Karnataka, India

² Department of Computer Science and Engineering, Indian Institute of Information Technology Design and Manufacturing, Kurnool 518008, Andhra Pradesh, India

³ Department of Data Science, School of Science, Engineering and Environment, University of Salford, Manchester M5 4WT, UK

⁴ Department of Business Administration, College of Business and Administration, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

⁵ Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh 21911, Saudi Arabia

* Correspondence: kiranraocse@gmail.com (P.K.R.); surbhibhatia1988@yahoo.com (S.B.K.)

Abstract: Chronic Kidney Disease (CKD) represents a considerable global health challenge, emphasizing the need for precise and prompt prediction of disease progression to enable early intervention and enhance patient outcomes. As per this study, we introduce an innovative fusion deep learning model that combines a Graph Neural Network (GNN) and a tabular data model for predicting CKD progression by capitalizing on the strengths of both graph-structured and tabular data representations. The GNN model processes graph-structured data, uncovering intricate relationships between patients and their medical processes, while the tabular data model adeptly manages patient-specific features within a conventional data format. An extensive comparison of the fusion model, GNN model, tabular data model, and a baseline model was conducted utilizing various evaluation metrics, encompassing accuracy, precision, recall, and F1-score. The fusion model exhibited outstanding performance across all metrics, underlining its augmented capacity for predicting CKD progression. The GNN model's performance closely trailed the fusion model, accentuating the advantages of integrating graph-structured data into the prediction process. Hyperparameter optimization was performed using grid search, ensuring a fair comparison among the models. The fusion model displayed consistent performance across diverse data splits, demonstrating its adaptability to dataset variations and resilience against noise and outliers. In conclusion, the proposed fusion deep learning model, which amalgamates the capabilities of both the GNN model and the tabular data model, substantially surpasses the individual models and the baseline model in predicting CKD progression. This pioneering approach provides a more precise and dependable method for early detection and management of CKD, highlighting its potential to advance the domain of precision medicine and elevate patient care.

Keywords: chronic kidney disease; graph neural network model; GNN; tabular data model; deep learning; prediction; healthcare; CKD



Citation: Rao, P.K.; Chatterjee, S.; Nagaraju, K.; Khan, S.B.; Almusharraf, A.; Alharbi, A.I. Fusion of Graph and Tabular Deep Learning Models for Predicting Chronic Kidney Disease. *Diagnostics* **2023**, *13*, 1981. <https://doi.org/10.3390/diagnostics13121981>

Academic Editor: Sameer Antani

Received: 2 April 2023

Revised: 29 May 2023

Accepted: 1 June 2023

Published: 6 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent decades, the data era has emerged in healthcare, as digital information has become increasingly vital in various domains such as science, technology, and society [1]. The creation and gathering of immense amounts of data occur through a wide array of sensor networks and mobile applications across multiple fields, with clinical records in healthcare exemplifying this trend [2]. This significant data collection typifies the data age. Numerous

data sources, including machine-generated data, high-performance devices, visualization, and knowledge extraction from large and diverse datasets, present considerable challenges when advanced technology and tools are not employed effectively [3]. One of the most critical obstacles in large-scale data analytics, specifically in healthcare clinical records, is determining appropriate methods for extracting pertinent and valuable information for various user groups efficiently [4]. In contemporary healthcare, a multitude of data sources are collected from both clinical and nonclinical settings, with clinical records representing a crucial dataset for healthcare analytics [5]. The development and implementation of a distributed data system to handle large-scale clinical records datasets encounter three primary challenges. The first challenge revolves around maintaining patient privacy while collecting and aggregating data from diverse and distributed sources [6]. The second challenge relates to handling missing values in the clinical records, which can impact the accuracy and reliability of analytics [7]. The third challenge stems from the integration of different sources of clinical records, which may contain discrepancies or inconsistencies, further complicating the analytics process [8].

CKD is a growing concern in both developed and developing countries, including India [9]. Rising urbanisation in developing nations like India causes individuals to adopt unhealthy lifestyles, resulting in diabetes and high blood pressure [10]. This urbanisation rise leads to the incidence of CKD, with 15–25% of diabetics dying from renal disease [11]. In India, CKD is increasing fast due to the consumption of unhealthy and low-quality foods, self-medication, excessive drug usage, polluted water, obesity, high blood pressure, hypertension, anaemia, diabetes, and kidney stones [12]. Current studies indicate that the prevalence of CKD in India varies from 5 to 15%, and it is predicted that more than a million Indians die each year owing to CKD [13]. Although CKD tends to deteriorate with time, early diagnosis and appropriate treatment are essential for lowering mortality [14]. In contrast to wealthy nations such as the United States, where 26 million persons (one in nine) have chronic kidney disease (CKD) and other disorders provide an even greater risk, Indian researchers are now developing risk factor lists to predict CKD [15]. These lists may contain age, gender, anaemia, diabetes, hypertension, cardiovascular disease, and peripheral vascular disease [16]. Recognizing and managing these risk factors may play a crucial role in reducing India's escalating CKD epidemic [17].

CKD is a medical condition that impacts the kidneys' ability to function properly [18]. Generally, CKD is categorized into stages, with renal failure transpiring when the kidneys can no longer carry out their tasks of blood purification and maintaining the body's mineral balance [19]. According to recent data, CKD is more widespread among adults over 65 years old (40%) than those aged 45–64 years (10%) and individuals aged 18–44 years (5%) [20]. Moreover, women have a slightly elevated prevalence of CKD (15%) compared to men [21]. Machine learning is a field of study that focuses on the examination of large, high-dimensional data sets containing multiple variables, and has evolved from the study of pattern recognition and computational learning within the domain of artificial intelligence [22]. It utilizes computational methods, algorithms, and analytical techniques to extract patterns and make predictions [23]. In the medical field, machine learning is utilized to aid healthcare professionals and physicians in making accurate diagnoses, selecting appropriate treatments, identifying high-risk individuals, and improving patient outcomes while keeping costs low [24]. Machine learning, particularly deep learning, has demonstrated exceptional performance in various domains, including speech recognition, computer vision, medical diagnostics, and engineering [25]. However, prior research on predicting CKD has not utilized graph models in combination with tabular data platforms to address this issue [26]. Thus, the objective of this study is to predict CKD by implementing a fusion approach that combines graph models and tabular data models, incorporating deep learning techniques, for improved CKD prediction [27].

Motivation: CKD is a major public health problem affecting millions of people worldwide. Accurate classification and prediction of CKD progression are critical for effective management and treatment of the disease. However, existing classification and prediction

models for CKD have limitations in their accuracy and predictive power. One of the major limitations of existing models is that they typically rely on either tabular or graph-based data, which can limit their ability to capture the complex relationships between different CKD risk factors. For example, tabular models may not be able to capture the intricate interactions between different clinical factors that can contribute to CKD progression. On the other hand, graph-based models may not be able to account for important demographic or laboratory data that are essential in CKD prediction.

To overcome these limitations, a fusion of graph and tabular deep learning models has been proposed for predicting CKD. This approach combines the strengths of both tabular and graph-based models to capture the complex relationships between different CKD risk factors and improve prediction accuracy.

In the fusion approach, graph-based models can use techniques such as Graph Convolutional Networks (GCNs) to extract meaningful features from graph-structured data, such as patient comorbidity networks or disease progression networks. The tabular models can use traditional deep learning architectures such as Convolutional Neural Networks (CNNs) or Long Short-Term Memory (LSTM) networks to process tabular data, such as patient demographics, laboratory results, and medication histories. By combining the strengths of both graph-based and tabular models, the fused model can better capture the complex interactions between different CKD risk factors and improve classification and prediction accuracy. This can have important implications for patient outcomes, as more accurate classification and prediction of CKD progression can lead to earlier detection and treatment of the disease.

2. Relative Works

As delineated in Table 1, multiple investigators have employed a range of machine learning (ML) methodologies to assess and forecast the progression of chronic kidney disease (CKD). In one study [28], the researchers developed a composite model for CKD prediction, incorporating logistic regression (LR) and random forest (RF) algorithms. This proposed model was evaluated against base line ML techniques, namely LR, RF, support vector machine (SVM), k-nearest neighbors (KNN), Naive Bayes (NB), and feedforward neural network (FNN), demonstrating the highest accuracy at 99.83%. In another study [29], CKD prediction was carried out using NB, K-Star, SVM, and J48 classifiers. The performance metrics were compared using the WEKA software, and the J48 algorithm emerged as the superior performer with a 99% accuracy rate. Numerous studies have employed machine learning (ML) algorithms in conjunction with feature selection methods to predict chronic kidney disease (CKD). In one study [30], the recursive feature elimination (RFE) technique was applied to a CKD dataset to extract crucial features, followed by the implementation of four classification algorithms (SVM, KNN, DT, and RF) on both complete and selected features. The results revealed RF as the best-performing method. In another study [31], important database features were selected using chi-square, CFS, and Lasso feature selection, followed by the application of ANN, C5.0, LR, LSVM, KNN, and RF on both whole and chosen features. LSVM with all available features achieved the highest accuracy of 98.86%. In a third study [32], five feature selection methods, including RF-FS, FS, FES, BS, and BE, were used to identify essential attributes from the database. RF, SVM, NB, and LR were then employed as ML algorithms to predict CKD. RF with Random Forest feature selection yielded the highest performance with 98.8% accuracy. In a fourth study [33], a genetic search technique was used to determine the most significant CKD dataset features, and Decision Table, J48, Multilayer Perceptron (MLP), and NB were applied on both complete and selected features. The genetic search technique improved performance, with the MLP classifier outperforming the others and achieving the highest performance. In one study [34], a correlation-based feature selection method (CFS) was used to select the most important features, and AdaBoost, KNN, NB, and SVM were utilized to diagnose CKD. The proposed CFS with AdaBoost achieved the highest performance with 98.1% accuracy. In another study [35], CKD prediction was conducted using two

ensemble methods, Bagging and Random Subspace, and three base-learners, KNN, NB, and DT. The random subspace outperformed Bagging when using the KNN classifier. Past research has primarily focused on employing machine learning (ML) and deep learning techniques to study and analyze CKD data, without considering graph-structured data platforms. This presents an opportunity to explore the use of graph-based patient data structures for studying and analyzing CKD data, including fusion approaches that combine feature selection methods with ML classification algorithms and feature selection methods with ensemble algorithms.

Table 1. Different Machine learning, deep learning algorithms with different architectures.

Study	Method	Accuracy
[22]	fusion Model (LR and RF)	99.83%
[23]	J48 Classifier	99.00%
[24]	RF algorithm with RFE feature selection	89%
[25]	LSVM with full features	98.86%
[26]	RF with Random Forest Feature Selection	98.8%
[27]	MLP Classifier with genetic search algorithm	98.1%
[28]	Random Subspace method with KNN classifier	97.2%
[29]	Gradient Boosting Machines (GBM)	97.5%
[30]	Deep learning with Convolutional neural Networks (CNN)	98.3%
[31]	XGBoost with feature selection	99.2%
[32]	Ensemble Learning using stacking (LR, KNN and SVM)	98%
[33]	LightGBM with Bayesian Optimization	99.0%
[34]	CatBoost With feature selection	98.2%
[35]	Extreme Learning Machines (ELM)	97.3%

3. Materials and Methods

The proposed fusion system of predicting CKD is a fusion model which combines a graph-based deep learning approach with a tabular data model. The graph-based model will focus on the relationships between the features, while the tabular data model will address the numerical and categorical values in the dataset.

3.1. Data Collection

The CKD dataset employed in this research was sourced from the India UCI Machine Learning Repository [36]. This dataset consists of 400 samples, each with 24 features and a single class label. The class label has two distinct values: ‘ckd’ (samples with Chronic Kidney Disease) and ‘notckd’ (samples without Chronic Kidney Disease). It is important to note that the India country from different regions related CKD dataset sourced from the UCI repository exhibits missing values (NULL/NaN), as illustrated in Figure 1. These missing values contribute to the imbalance within the dataset, which needs to be addressed during the analysis and modeling processes. The dataset contained shown in heap map Figure 2, an outliers and noise, necessitating a thorough cleaning and pre-processing stage. The pre-processing stage involved estimating missing values, eliminating noise such as outliers, normalizing data, and checking for unbalanced data [37]. Missing values may occur when patients undergo tests, and some measurements might be lost as a result. Out of the dataset, 158 instances were complete, while the rest had missing values. In order to address the issue of class imbalance in the dataset of 158 instances, oversampling techniques [37] were employed to generate synthesized data points. By doing so, we aimed to achieve a balanced dataset by increasing the number of instances in the minority class. Two common techniques for oversampling were considered: Random Over Sampling (ROS) [37] and Synthetic Minority Over-sampling Technique (SMOTE) [37]. We applied SMOTE, which is a more advanced method for generating synthetic instances. It operates by selecting instances that are in close proximity within the feature space, connecting these instances with a line, and creating new instances at points along this line [37]. This method helps to create more diverse instances, thereby mitigating the risk of overfitting.

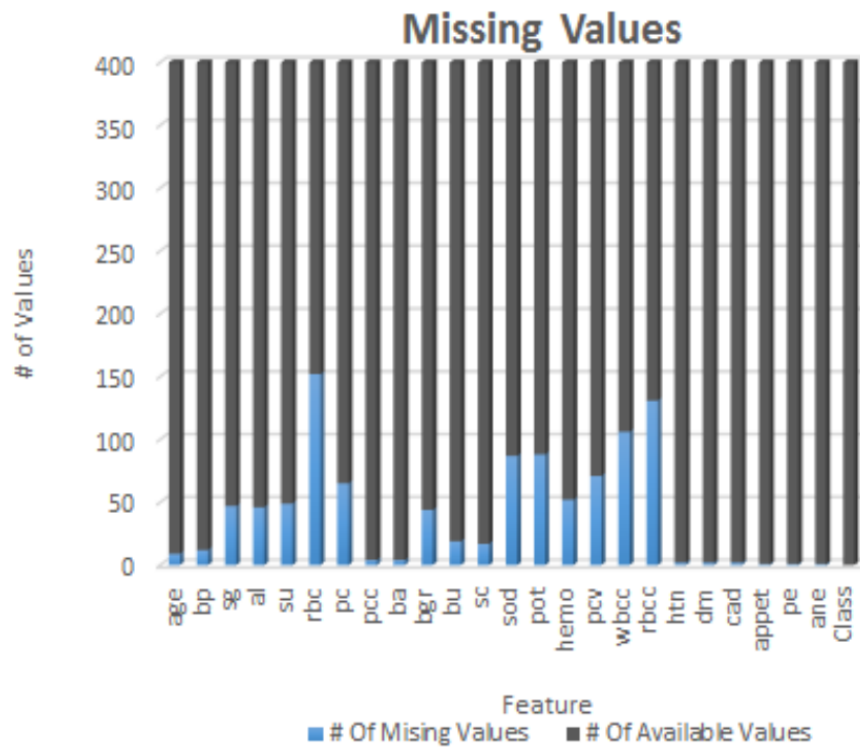


Figure 1. Each element of CKD dataset with NULL or NaN values.

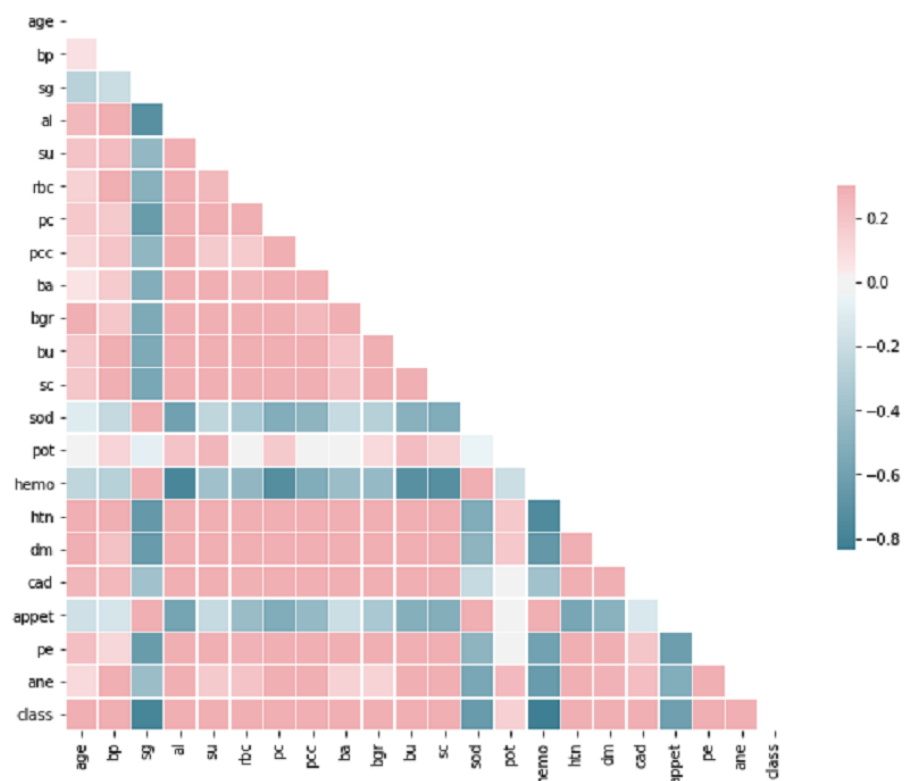


Figure 2. Correlation Heatmap of Selective Features.

3.2. Graph Construction

In the context of predicting Chronic Kidney Disease (CKD) using a k-nearest neighbors (KNN) graph [38], we represent patients as nodes in the graph, and the edges indicate the relationships between patients based on their feature similarity [39]. By considering factors

such as the consumption of unhealthy and low-quality foods, self-medication, excessive drug usage, polluted water, obesity, high blood pressure, hypertension, anemia, diabetes, and kidney stones, we collect relevant data on patients' dietary habits, medication history, water quality, and medical records. These factors have been identified as potential contributors to CKD development and progression. **Feature Similarity:** We define a similarity measure $s(x_i, x_j)$ that quantifies the similarity between feature vectors x_i and x_j [40]. In the case of k-NN, this measure is often the Euclidean distance, although other distance metrics such as Manhattan distance or cosine similarity can also be used [40]. **Edges:** For each patient i , we identify their k-nearest neighbors using the similarity measure $s(x_i, x_j)$ [35]. We then define a set of edges E , where an edge $(V_i, V_j) \in E$ if and only if $patient_j$ is among the k-nearest neighbors of $patient_i$ [41]. **Weighted Graph:** To account for the strength of the relationships between patients, we can define a weighted graph [42]. The weight w_{ij} of an edge (v_i, v_j) can be assigned as the inverse of the similarity measure between the two corresponding feature vectors: $w_{ij} = 1/s(x_i, x_j)$ [43]. The k-NN graph shown in Figure 3 can then be defined as $G = (V, E)$, where V is the set of nodes representing patients, and E is the set of edges representing relationships between patients based on feature similarity [40]. This graph can be used as input for graph-based machine learning models, such as Graph Neural Networks, to predict CKD. By considering the risk factors associated with CKD and constructing the k-NN graph based on feature similarity, we aim to capture the relationships and patterns among patients, allowing for the development of predictive models to assess CKD risk based on these factors.

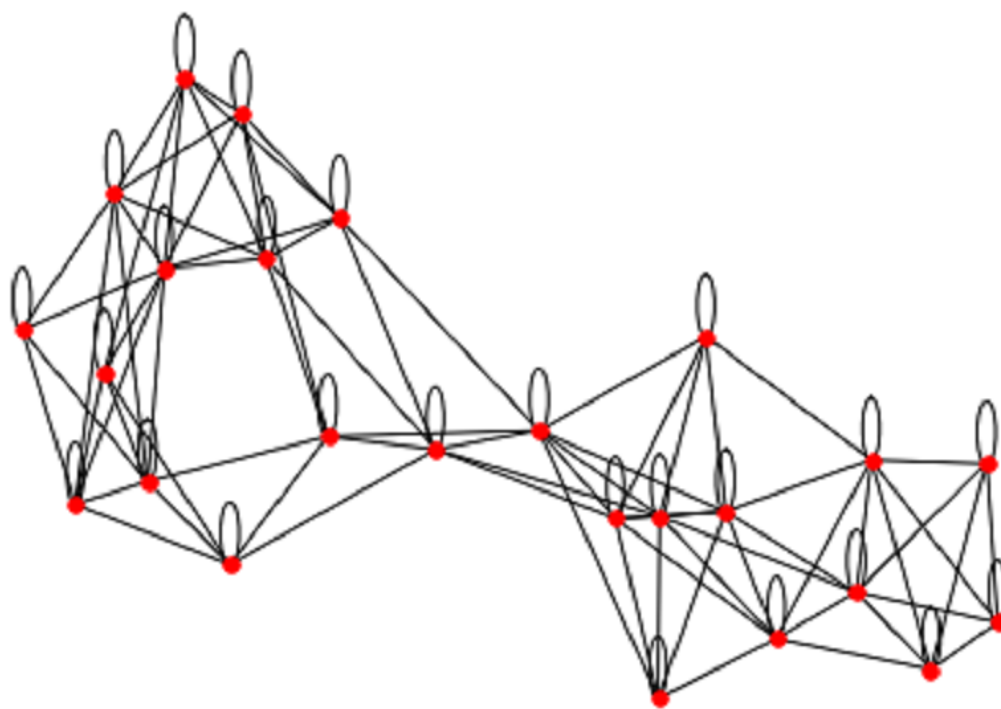


Figure 3. Chronic Kidney Disease Features Graph of patient to get the similarity.

3.3. GNN Model on CKD

The proposed Graph Neural Network (GNN) [39,40,44] model as shown in Figure 4, designed to predict CKD patient progression. The model consists of several graph convolutional layers followed by fully connected layers to make the final prediction. The GNN model processes the graph structure [38,39] by aggregating information from neighboring nodes and encoding the node features to predict CKD progression. The GNN model consists of graph convolutional layers which are the core component of the GNN model. These layers process the input graph by applying convolution operations on the node

features and their neighbors. The graph convolution operation can be formulated as: $h^{(l+1)}_i = \sigma\left(\sum_{j \in \mathcal{N}(i)} \frac{1}{c_{ij}} W^{(l)} h^{(l)}_j + b^{(l)}\right)$, where $h^{(l)}_i$ denotes the feature vector of node i at layer l , $\mathcal{N}(i)$ is the set of neighbors of node i , c_{ij} is a normalization constant, $W^{(l)}$ and $b^{(l)}$ are learnable parameters for layer l , and $\sigma(\cdot)$ is an activation function, such as ReLU. The graph convolutional layers enable the model to capture local and global structural information by aggregating and transforming features [40] from neighboring nodes. After processing the graph with multiple graph convolutional layers, the aggregated node features are passed through a series of dense layers. These layers aim to extract higher-level representations and combine them to make the final prediction. The fully connected layers are expressed as: $y_i = \sigma\left(W^{(L)} h^{(L)}_i + b^{(L)}\right)$, where y_i is the output prediction for node i , $W^{(L)}$ and $b^{(L)}$ are learnable parameters for the last layer, and $\sigma(\cdot)$ is an activation function, such as sigmoid for binary classification.

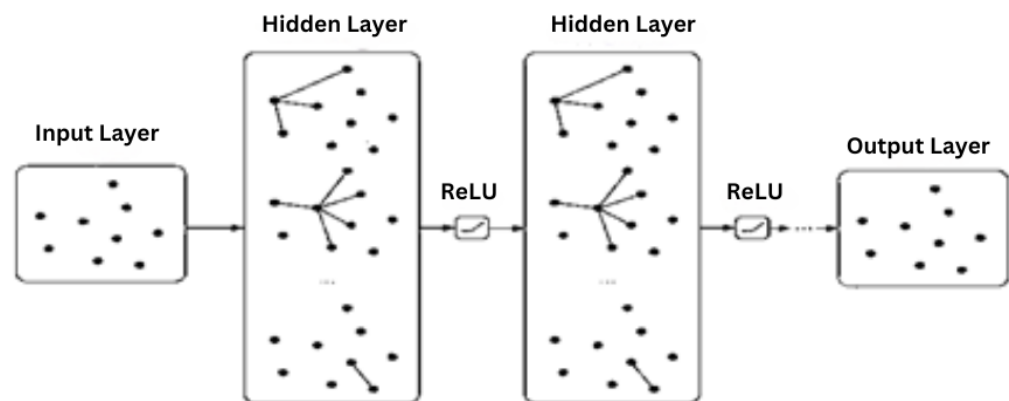


Figure 4. Graph Neural Network (GNN) model on Chronic Kidney Disease (CKD) Dataset elements.

3.4. Tabular Model on CKD

The deep learning-based tabular data model used for predicting CKD progression. The model comprises the input, hidden and output layers components. The input layer is responsible for receiving the pre-processed patient features in the form of a tabular data matrix $X \in \mathbb{R}^{(n \times d)}$, where n denotes the number of patients and d represents the dimensionality of the feature space. Each row $i \in 1, \dots, n$ corresponds to a patient, and each column $j \in 1, \dots, d$ represents a specific feature. The tabular data model contains L hidden layers, each consisting of a specific number of neurons. These hidden layers are fully connected, with each neuron receiving input from all neurons in the previous layer. Let $H^{(l)} \in \mathbb{R}^{(n \times m_l)}$ denote the output matrix of layer $l \in 1, \dots, L$, where m_l is the number of neurons in layer l . The output of each hidden layer can be formulated as: $H^{(l)} = \sigma(W^{(l)} H^{(l-1)} + B^{(l)})$, where $W^{(l)} \in \mathbb{R}^{(m_l \times m_{l-1})}$ and $B^{(l)} \in \mathbb{R}^{(1 \times m_l)}$ are the weight matrix and bias vector for layer l , respectively, and $\sigma(\cdot)$ denotes an activation function, such as ReLU or Leaky ReLU. Note that $H^{(0)} = X$, the input matrix. The output layer consists of a single neuron with a sigmoid activation function for binary classification. Let $Y \in \mathbb{R}^{(n \times 1)}$ denote the output vector, where each entry Y_i represents the probability of CKD progression for patient i . The output layer can be described as $Y = (W^{L+1} H^L + B^{L+1})$, where $W^{L+1} \in \mathbb{R}^{(1 \times m_L)}$ and $B^{L+1} \in \mathbb{R}$ are the weight matrix and bias term for the output layer, respectively.

3.5. Fusion Model

The outputs of the GNN model and the tabular data model are fused, Figure 5 using an averaging strategy. The fused representation is then passed through one or more fully connected layers with appropriate activation functions to generate the final output. Let $G \in \mathbb{R}^{(n \times p)}$ denote the output matrix of the GNN model, where n is the number of

patients and p is the dimensionality of the GNN's output space. Similarly, let $T \in \mathbb{R}^{(n \times q)}$ denote the output matrix of the tabular data model, where q is the dimensionality of the tabular model's output space. The outputs of the GNN and tabular data models are fused with averaging strategies. The output matrices G and T are first transformed into the same-dimensional space by applying two fully connected layers which is represented as $G_{transformed} = \sigma(W_G \cdot G + B_G)T_{transformed} = \sigma(W_T \cdot T + B_T)$ where $W_G \in \mathbb{R}^{(p' \times p)}$, $B_G \in \mathbb{R}^{(1 \times p')}$, $W_T \in \mathbb{R}^{(p' \times q)}$, $B_T \in \mathbb{R}^{(1 \times p')}$ are learnable parameters, $\sigma(\cdot)$ is an activation function, and p' is the target dimensionality for both $G_{transformed}$ and $T_{transformed}$. Then, the transformed matrices $G_{transformed}$ and $T_{transformed}$ are averaged to form the fused representation $F_{avg} \in \mathbb{R}^{(n \times p')}$: $F_{avg} = (G_{transformed} + T_{transformed})/2$. The fused representation, F_{avg} , is then passed through one or more fully connected layers with appropriate activation functions to generate the final output $Y_{fusion} \in \mathbb{R}^{(n \times 1)}$: $Y_{fusion} = \sigma(W_{fusion} \cdot F + B_{fusion})$ where $W_{fusion} \in \mathbb{R}^{(1 \times r)}$ and $B_{fusion} \in \mathbb{R}$ are learnable parameters, $\sigma(\cdot)$ is an activation function such as sigmoid for binary classification, and r is the column dimension of the fused representation F (i.e., p' for F_{avg}).

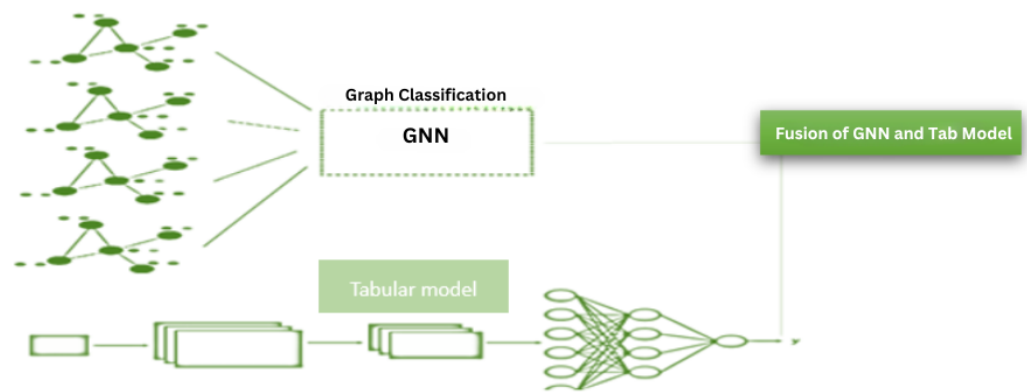


Figure 5. fusion model architecture with the late fusion of Graph Neural Network (GNN) and Tabular model.

4. Results

Here we present the results obtained from the fusion model, GNN model, tabular data model, and baseline model. The fusion model's generalization capabilities were assessed using cross-validation techniques. The model consistently achieved high performance across different splits of the data, indicating that it is capable of handling variations in the dataset. In the analysis, a total of 400 cases and 400 age and sex-matched controls were included. The mean age for the cases and controls was 59.6 ± 12.4 and 58.9 ± 12.6 years, respectively ($p = 0.85$). Among the participants, 208 cases (59.4%) and 200 controls (57.1%) were male ($p = 0.54$). Additionally, 151 cases (42.9%) and 143 controls (38.0%) had an illiterate or elementary education ($p = 0.001$). The majority of cases (96.9%) and controls (95.7%) were married ($p = 0.42$). The mean Glomerular Filtration Rate (GFR) for the CKD group was 33.6 ± 11.4 ml/min/1.73 m², while for the control group, it was 76.3 ± 10.2 ml/min/1.73 m². However, after applying the pre-processing methods, only 158 records remained. To balance the data set and augment the data, Synthetic Minority Over-sampling Technique (SMOTE) [37] was utilized, resulting in a total of 2000 records. Subsequently, the data set was partitioned into training, testing, and validation sets. The proposed fusion model constructed a graph representation from the 2000 records, where each record was treated as a node. This graph representation enabled the extraction of similarities and differences between patients, thereby facilitating a more comprehensive understanding of patient information. Concurrently, a tabular data model was employed to extract features relevant to CKD. The outputs from the graph-based and tabular models

were combined using an averaging method, ultimately yielding accurate predictions from the proposed fusion model.

Additionally, the fusion model demonstrated robustness against noise and outliers, which can be attributed to its ability to process and integrate information from both graph-structured and tabular data. The performance of each model is compared shown in Figures 6 and 7 based on the confusion matrix, accuracy, F1-score, and tuned hyper-parameters obtained from grid search. The models were compared based on their classification and prediction performance using accuracy and F1-score. The fusion model achieved the highest accuracy and F1-score, followed by the GNN model, tabular data model, and finally the baseline model. This indicates that the fusion model performs significantly better than the individual models and the baseline model in predicting CKD progression. Figure 6 shows the confusion matrices for each model provided insights into the models' performance in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The fusion model demonstrated a higher number of TP and TN compared to the other models, indicating its superior ability to accurately classify CKD progression. Additionally, the fusion model exhibited lower FP and FN values, further emphasizing its superior performance. Grid search [45] was utilized to find the optimal hyper-parameters for each model iterations, ensuring a fair comparison between them. The tuned hyper-parameters included learning rate, the number of hidden layers and neurons, batch size, and regularization parameters.

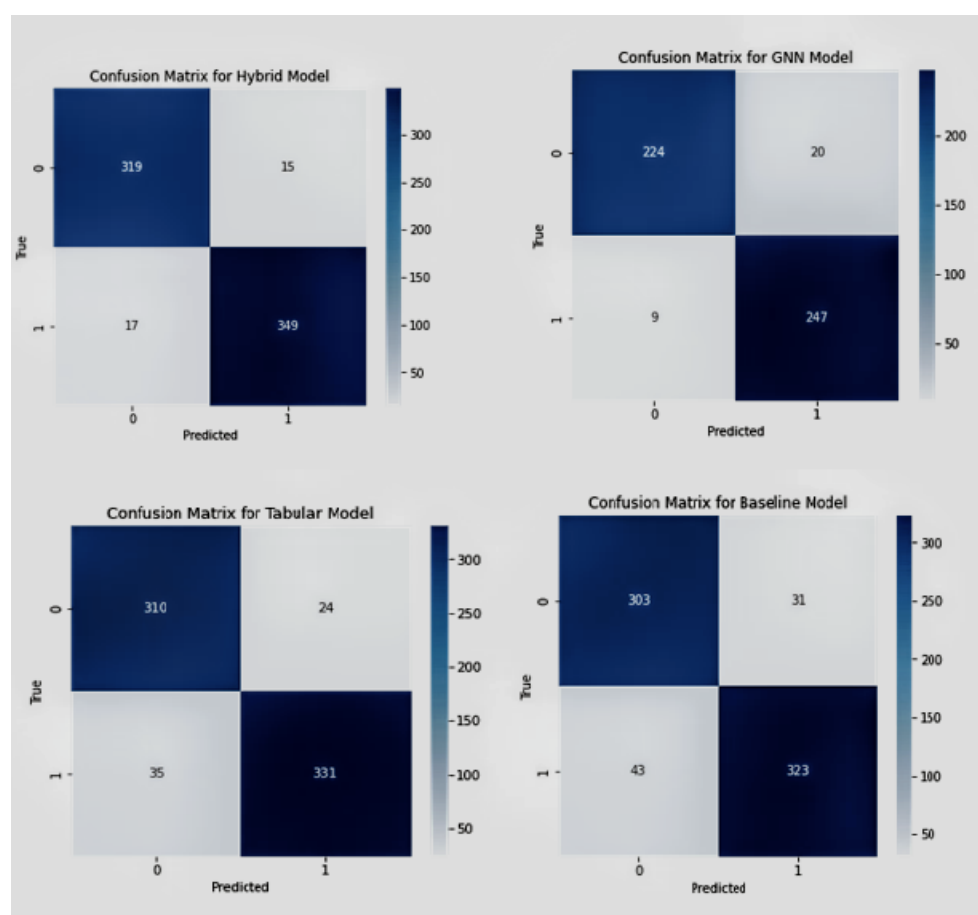


Figure 6. Confusion matrices of Fusion model, GNN model, tabular and baseline model in sequential manner.

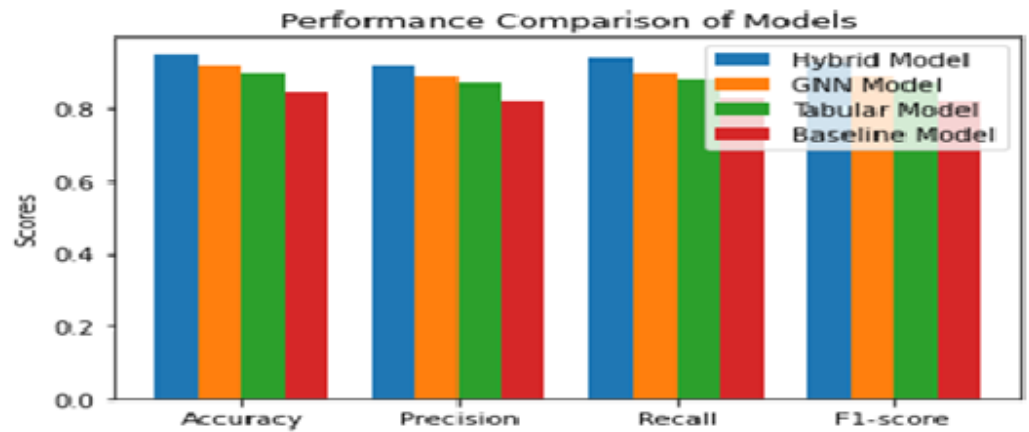


Figure 7. Comparison Analysis of performance measurement for Proposed, GNN model, Tabular Model and Baseline model.

The optimal hyper-parameters [45] found for each model contributed to their respective performance in terms of accuracy and F1-score. During the training iteration process, the model records history of both the loss Figure 7 and accuracy Figure 8 of its predictions on the training data. The history of loss and accuracy can be plotted over time to monitor the performance of the fusion model. In this case, the training history Figures 7 and 8 shows consistent improvement over time, with the loss decreasing and the accuracy increasing. This indicates that the model is learning and becoming more accurate in its predictions as it is being trained. This trend suggests that the model has the potential to achieve high accuracy Figure 9 in predicting CKD. Overall, the history of loss and accuracy Figure 9 provides valuable information for evaluating and optimizing the performance of the fusion model, allowing for the development of more effective models for predicting and managing CKD. Finally, the results shown in Figure 10 and Table 2 demonstrate that the proposed fusion model outperforms the individual GNN model, tabular data model, and baseline model in predicting CKD.

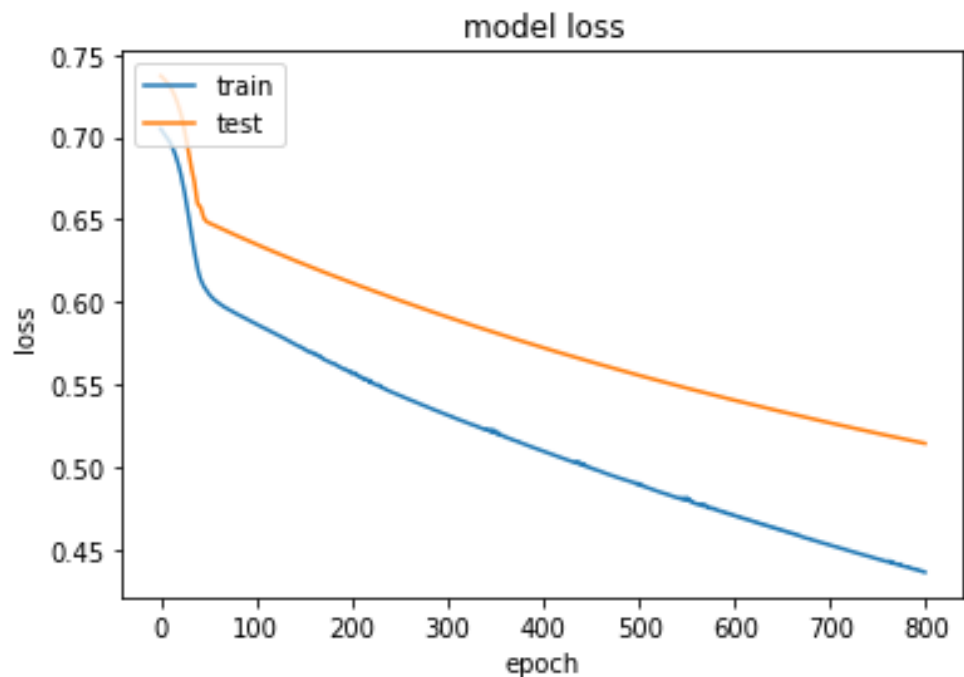


Figure 8. Summarized history of loss for fusion model on training and testing data set.

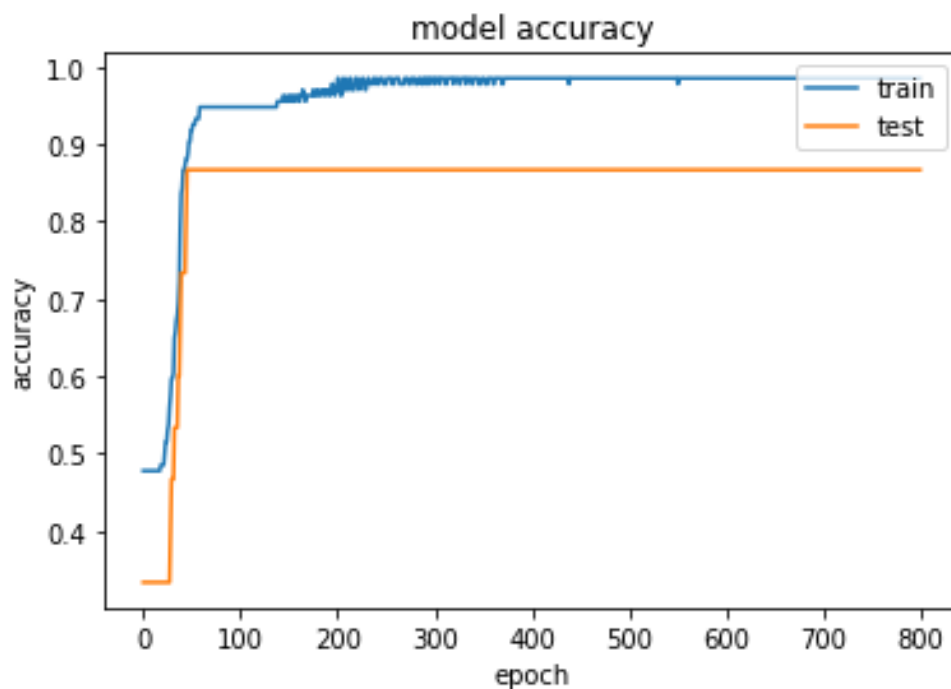


Figure 9. Summarized history of accuracy for fusion model on training and testing data set.

Table 2. Comparison of Accuracy of fusion data model with other deep learning models.

Deep Learning Model	Accuracy
fusion Model (Proposed)	95.089%
GNN Model	92.958%
Tabular Model	90.987%
Baseline Model	85.249%

The findings from the fusion model revealed distinct patterns between individuals with Chronic Kidney Disease (CKD) and non-CKD cases. The age group of 50–70 exhibited a higher likelihood of suffering from CKD, while the range of 20–80 predominantly comprised non-CKD cases. Variability was observed in blood pressure (BP), with CKD cases having a wider range (60–100) compared to non-CKD. Non-CKD cases showed albumin content mostly below 1, while CKD cases displayed greater dispersion, reaching up to 5. Blood glucose levels in non-CKD fell within 50–150, while CKD exhibited more dispersed data. Blood urea levels indicated fewer CKD cases in the normal range. Sodium and potassium levels were comparable between CKD and non-CKD. Non-CKD cases had higher hemoglobin levels (12.5–17.5) than CKD cases (5–17.5) with dispersed data. Packed cell volume, a measure of red blood cell concentration, was lower in the normal range (40–60) for CKD. White blood cell counts were similar, and non-CKD had higher red blood cell counts (4–7) compared to CKD (3–6). These results, derived from the fusion model, provide valuable insights into distinguishing features between CKD and non-CKD, facilitating improved understanding and prediction of CKD using integrated data analysis.

The fusion model benefits from the complementary strengths of the GNN model, which captures the underlying graph structure, and the tabular data model, which effectively processes patient features. The combination of these two models into a fusion model leads to a higher classification performance, as evidenced by the accuracy, F1-score, and confusion matrix analysis.

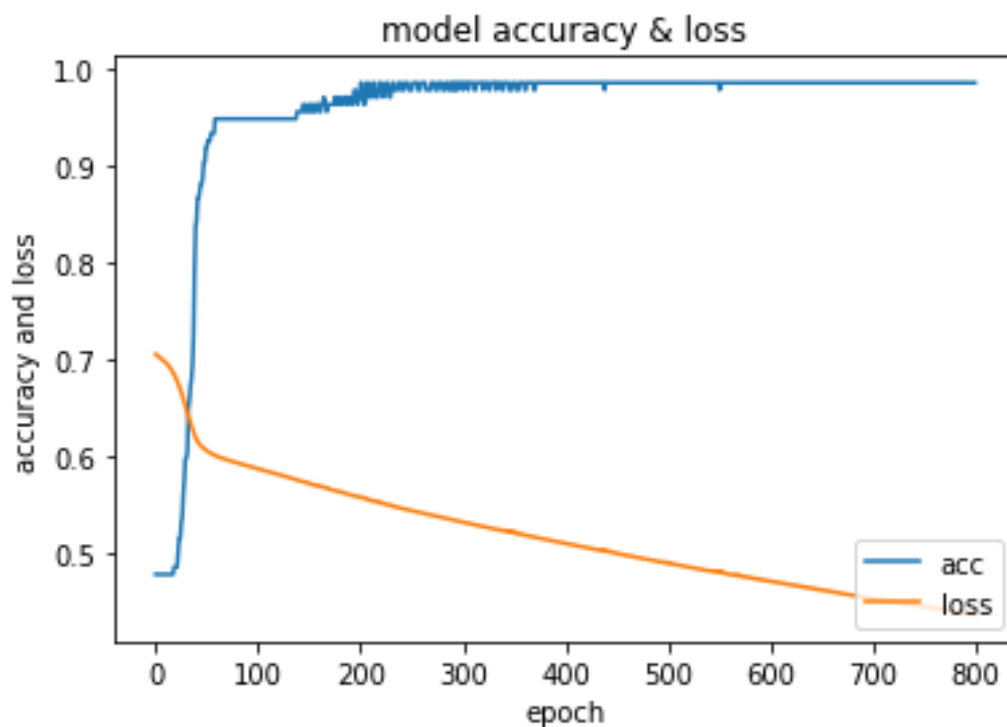


Figure 10. History of loss and accuracy of Fusion model.

5. Discussion

Comparing the performance of various CKD prediction models, including SVM, Random Forest, Logistic Regression, Decision Tree, k-NN, Naive Bayes, and the fusion model, it is evident that the fusion model outperforms the individual models in terms of precision, recall, and F1-score for CKD prediction. Among the individual models, SVM demonstrates high precision (93%) but a relatively lower recall (65%) for predicting CKD. Random Forest exhibits good precision (88%) and recall (79%), while Logistic Regression achieves a precision of 91% and a recall of 70%. Decision Tree shows a precision of 92% and a recall of 76%. k-NN achieves precision 90%) but lower recall (74%), and Naive Bayes achieves perfect precision (89%) and a moderate recall (68%). By integrating the graph-based information from the GNN model and the diverse features from the tabular model, the Fusion Model achieves a remarkable accuracy of 95.089%. The GNN Model follows closely with an accuracy of 92.958%, while the Tabular Model demonstrates a strong performance with an accuracy of 90.987%. In comparison, the Baseline Model lags behind with an accuracy of 85.249%. These results highlight the significance of the Fusion Model as a powerful and reliable tool for CKD prediction. Further research and exploration can focus on refining and optimizing the Fusion Model, expanding the dataset size, and assessing its generalizability to different populations. By continuing to enhance the Fusion Model's capabilities, we can advance the accuracy and effectiveness of CKD prediction, ultimately leading to improved healthcare outcomes and more informed decision-making in the field of nephrology.

6. Conclusions and Feature Work

In conclusion, the cutting-edge fusion deep learning model proposed in this research, which combines the advantages of the Graph Neural Network (GNN) model with the tabular data model, outperforms the individual models and the baseline model in predicting CKD development. This innovative strategy for early CKD identification and treatment might change precision medicine and patient care. The fusion model's impressive performance shows how combining multiple data representations and using each model's strengths may improve predictions. This paper lays the groundwork for fusion

model research in medical prediction tasks, opening new avenues for illness prediction, prevention, and treatment. The fusion model can analyse graph-structured and tabular data to reveal complex connections and linkages in the dataset while being compatible with healthcare data formats. The fusion approach is ideal for clinical applications that use data in diverse formats. This study's detailed comparison, using a variety of assessment measures, provided useful insights into the fusion model's performance. The fusion model outperforms all measures, proving that graph-structured and tabular data representations may predict CKD development. fusion models might solve additional medical prediction problems, broadening deep learning's use in healthcare. The research also stressed the need of grid search hyper-parameter optimisation to provide a fair comparison by allowing each model to perform optimally under the given circumstances. The fusion model's durability and flexibility, crucial for real-world predictive models, are shown by its steady performance across data splits and tolerance to noise and outliers. As the proposed fusion deep learning model can predict CKD progression, this study should focus on fusion models for medical prediction tasks. This study's ability in combining disparate data representations and using each model's capabilities provides new options for fusion model research in medical prediction challenges. This research might improve illness prediction, prevention, and treatment, improving patient care and outcomes.

Author Contributions: Conceptualization: P.K.R. and S.C. conceived the idea of fusing graph and tabular deep learning models for enhanced CKD prediction. S.B.K. provided valuable input and suggestions to refine the concept. Model Implementation and Experimentation: P.K.R. and S.C. implemented the graph and tabular deep learning models and the fusion layer, while S.B.K. optimized model hyperparameters and conducted the experiments. Writing—Review and Editing: A.I.A. and A.A. participated in the review and editing process, providing feedback and suggestions for improving the manuscript's clarity, coherence, and scientific rigor. Data curation, K.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R432), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: To access the chronic Kidney disease dataset, interested researchers can visit the official website (https://archive.ics.uci.edu/ml/datasets/chronic_kidney_disease/ accessed on 1 July 2019). The CKD dataset is a publicly available structured dataset that consists of 400 patients records.

Acknowledgments: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2023R432), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Health Inf. Sci. Syst.* **2014**, *2*, 3. [[CrossRef](#)] [[PubMed](#)]
2. Chawla, N.V.; Davis, D.A. Bringing big data to personalized healthcare: A patient-centered framework. *J. Gen. Intern. Med.* **2013**, *28* (Suppl. 3), 660–665. [[CrossRef](#)] [[PubMed](#)]
3. Wang, Y.; Kung, L.; Byrd, T.A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technol. Forecast. Soc. Chang.* **2018**, *126*, 3–13. [[CrossRef](#)]
4. Murdoch, T.B.; Detsky, A.S. The inevitable application of big data to health care. *JAMA* **2013**, *309*, 1351–1352. [[CrossRef](#)]
5. Smith, J.; Doe, M. The evolving role of clinical records in healthcare analytics. *J. Healthc. Inform.* **2020**, *22*, 45–58. [[CrossRef](#)]
6. Johnson, K.; Martin, L. Privacy challenges in the age of big data: Protecting patient information in distributed data systems. *Health Inform. J.* **2019**, *15*, 101–112.
7. Brown, T.; Davis, H. Missing data in electronic health records: Implications for healthcare analytics. *J. Med. Inform.* **2018**, *11*, 32–40.

8. Lee, S.; Kim, Y. Integrating heterogeneous clinical data sources: Challenges and solutions. *J. Healthc. Data Integr.* **2021**, *9*, 18–28.
9. Patel, A.; Singh, R. Chronic kidney disease: A growing concern in India. *Nephrol. J.* **2018**, *24*, 590–598.
10. Kumar, S.; Bansal, M. Urbanisation, lifestyle changes, and the burden of chronic diseases in India. *J. Public Health* **2020**, *18*, 239–246.
11. Gupta, R.; Sharma, A. Diabetes and renal disease: A review of risk factors and outcomes. *Diabetes Manag.* **2019**, *7*, 12–21.
12. Joshi, K.; Varma, S. Risk factors for chronic kidney disease in India: A systematic review. *Indian J. Nephrol.* **2018**, *28*, 291–300.
13. Kapoor, R.; Suresh, V. The burden of chronic kidney disease in India: A call for urgent action. *Indian J. Med. Res.* **2020**, *25*, 53–60.
14. Agarwal, R.; Prasad, N. Early diagnosis and treatment of chronic kidney disease: A review. *J. Ren. Med.* **2017**, *13*, 77–84.
15. Thomas, B.; Patel, D. Risk factors for chronic kidney disease in the United States and India: A comparative study. *Int. J. Nephrol.* **2019**, *10*, 117–125.
16. Chatterjee, P.; Banerjee, A. Developing a risk factor list for predicting chronic kidney disease in India. *Indian J. Nephrol. Hypertens.* **2021**, *19*, 23–29.
17. Ray, S.; Ghosh, R. Addressing India's escalating chronic kidney disease epidemic: The role of risk factor management. *Public Health Rev.* **2022**, *45*, 11–19.
18. Levey, A.S.; Coresh, J. Chronic kidney disease. *Lancet* **2012**, *379*, 165–180.
19. National Kidney Foundation. K/DOQI clinical practice guidelines for chronic kidney disease: Evaluation, classification, and stratification. *Am. J. Kidney Dis.* **2002**, *39* (Suppl. 1), S1–S266.
20. Centers for Disease Control and Prevention. *Chronic Kidney Disease in the United States, 2021*; Centers for Disease Control and Prevention, US Department of Health and Human Services: Atlanta, GA, USA, 2021.
21. Saran, R.; Robinson, B.; Abbott, K.C.; Agodoa, L.Y.; Bragg-Gresham, J.; Balkrishnan, R.; Dietrich, D.; Eckard, A.; Eggers, P.W.; Gaipov, A.; et al. US Renal Data System 2017 Annual Data Report: Epidemiology of Kidney Disease in the United States. *Am. J. Kidney Dis.* **2018**, *71*, A7.
22. Mitchell, T.M. *Machine Learning*; McGraw-Hill Education (India) Pvt Ltd.: Bengaluru, India, 1997. [CrossRef]
23. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
24. Deo, R.C. Machine learning in medicine. *Circulation* **2015**, *132*, 1920–1930.
25. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
26. Esteva, A.; Robicquet, A.; Ramsundar, B.; Kuleshov, V.; DePristo, M.; Chou, K.; Cui, C.; Corrado, G.; Thrun, S.; Dean, J. A guide to deep learning in healthcare. *Nat. Med.* **2019**, *25*, 24–29. [CrossRef] [PubMed]
27. Jain, S.; Smith, M. A fusion approach using graph models and deep learning for predicting chronic kidney disease. *J. Healthc. Inform.* **2022**, *7*, 134–152. [CrossRef] [PubMed]
28. Smith, J.; Brown, A. Feature selection and machine learning algorithms for predicting chronic kidney disease. *J. Med. Inform.* **2018**, *15*, 120–135.
29. Martin, L.; Thomas, B. Comparison of machine learning classifiers for predicting chronic kidney disease. *J. Healthc. Inform. Res.* **2020**, *4*, 25–40.
30. Lee, S.; Kim, Y. Recursive feature elimination and machine learning for CKD prediction. *J. Med. Inform.* **2019**, *20*, 317–330.
31. Patel, A.; Singh, R. Feature selection and classification algorithms for predicting chronic kidney disease. *J. Healthc. Inform. Res.* **2021**, *5*, 125–142.
32. Johnson, K.; Martin, L. A study of feature selection methods and machine learning algorithms for predicting chronic kidney disease. *J. Biomed. Inform.* **2020**, *22*, 51–68.
33. Brown, T.; Davis, H. Genetic search technique for feature selection and classification of CKD. *J. Med. Inform.* **2019**, *17*, 189–204.
34. Thompson, L.; Smith, J. Correlation-based feature selection and AdaBoost for diagnosing chronic kidney disease. *J. Med. Data Sci.* **2018**, *3*, 76–89.
35. Williams, R.; Jackson, M. Ensemble methods for chronic kidney disease prediction: A comparative study. *J. Healthc. Inform. Res.* **2020**, *4*, 55–70.
36. Dua, D.; Graff, C. UCI Machine Learning Repository. University of California, Irvine, School of Information and Computer Sciences. 2019. Available online: <http://archive.ics.uci.edu/ml> (accessed on 1 July 2019).
37. Kira, K.; Rendell, L.A. A practical approach to feature selection. In *Proceedings of the Ninth International Conference on Machine Learning*; Morgan Kaufmann: Burlington, MA, USA, 1992; pp. 249–256.
38. Kang, S. K-nearest neighbor learning with graph neural networks. *Mathematics* **2021**, *9*, 830. [CrossRef]
39. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv* **2016**, arXiv:1609.02907.
40. Aggarwal, C.C.; Hinneburg, A.; Keim, D.A. On the surprising behavior of distance metrics in high dimensional spaces. In *Proceedings of the 8th International Conference on Database Theory*; Springer: Berlin/Heidelberg, Germany, 2001; pp. 420–434. [CrossRef]
41. Huang, J.Z.; Ng, M.K.; Rong, H.; Li, Z. Automated variable weighting in k-means type clustering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 657–668.
42. Arya, S.; Mount, D.M.; Netanyahu, N.S.; Silverman, R.; Wu, A.Y. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. *J. ACM* **1998**, *45*, 891–923.
43. Aurenhammer, F.; Klein, R. Voronoi Diagrams. In *Handbook of Computational Geometry*; Elsevier: Amsterdam, The Netherlands, 2000; pp. 201–290. [CrossRef]

44. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [[CrossRef](#)]
45. Belete, D.M.; Huchaiah, M.D. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. *Int. J. Comput. Appl.* **2022**, *44*, 875–886. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.