Journal of the Postgraduate Institute of Medicine 2026; 13(1): E290 1-23

DOI: https://doi.org/10.4038/jpgim.v13i1.8599

Original Article

The Effect of Sparse Data Problem on Classification Algorithms in Machine Learning: A Case Study on Early-Stage Diabetes Risk Prediction

Thevaka Satheeskumar[1], Sampath Deegalla[2], Thirusittampalam Ketheesan[3]

[1]University of Vavuniya, [2]University of Peradeniya [3]University of Jaffna

Key words: Diabetes Mellitus, Early Diagnosis, Machine Learning, Diagnostic Techniques and Procedures

Abstract

Background and Objective/s

Machine learning (ML) is increasingly applied in healthcare to support early diagnosis of chronic diseases such as diabetes mellitus. However, its performance under sparse data conditions—where values are missing or incomplete—remains insufficiently understood, despite such conditions being common in patient-reported outcomes and non-clinical assessments. This study evaluates the performance of four supervised classification algorithms—Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and Naive Bayes (NB)—in predicting early-stage diabetes risk from a sparse dataset obtained through self-reported questionnaires.

Methods

A comparative experimental design assessed classifiers on a healthcare dataset of 677 records from individuals at risk of early-stage diabetes mellitus, collected from diabetic clinics and outpatient departments at Teaching Hospital, Jaffna. Data sparsity (10–40%) was introduced synthetically at controlled levels using a zero-placeholder encoding strategy to model patient perceptual uncertainty, while preserving core clinical features. Hyperparameter tuning was performed using randomized search with stratified 10-fold cross-validation. Performance was evaluated using accuracy, precision, recall, F1-score, sensitivity, specificity, and ROC–AUC, with statistical significance testing (paired t-test, Wilcoxon signed-rank test, Friedman test) to validate comparative findings. Memory efficiency was assessed by comparing standard and sparse matrix storage formats.

Results

RF demonstrated superior robustness across all sparsity levels, maintaining accuracy above 0.87 and AUC above 0.91 up to 40% sparsity. DT showed moderate resilience (AUC: 0.85 at 40% sparsity), while SVM and NB proved highly sensitive to increasing uncertainty, with AUC declining to 0.62 and 0.55 respectively at 40% sparsity. Statistical significance testing confirmed that performance differences between RF and other classifiers were significant ($p < 0.05$) with large effect sizes (Cohen's $d > 0.8$). Sparse matrix formats reduced memory usage by approximately 62% without compromising predictive accuracy.

Conclusions/Recommendations

RF emerged as the most reliable model for early-stage diabetes risk prediction under conditions of patient perceptual uncertainty, demonstrating that ensemble-based methods can accommodate subjective ambiguity inherent in self-reported symptom data. The zero-placeholder encoding strategy offers a transparent and interpretable approach to representing

https://doi.org/10.4038/jpgim.8599

uncertainty in ordinal clinical variables. While external validation is required, these findings provide a proof-of-concept that sparsity-aware ML models can support scalable digital health screening, particularly when conventional diagnostic techniques are unavailable. Integrating such approaches with uncertainty-aware preprocessing may enhance early detection and facilitate personalized healthcare delivery.

Corresponding Author: Thevaka Satheeskumar, E-mail thevaka_sathees@vau.ac.lk >

https://orcid.org/0009-0008-1058-5387

Received: 15 Sep 2025, Accepted revised version: 20 Mar 2026, Published: 23 Mar 2026

Competing Interests: Authors have declared that no competing interests exist

Introduction

Machine learning (ML) has emerged as a transformative tool in healthcare, enabling advanced diagnostic techniques and procedures for the early diagnosis of chronic diseases such as diabetes mellitus [1,2]. By leveraging algorithms and statistical models, ML systems can analyze complex datasets and identify subtle patterns that may not be evident through conventional statistical approaches [3]. ML algorithms are designed to learn from historical data, continuously improving their predictive accuracy without the need for explicit reprogramming [4]. These algorithms are commonly categorized into supervised, unsupervised, and reinforcement learning [5], each tailored to specific problem types and outcome objectives.

In supervised learning, which is the focus of this study, models learn from labeled datasets to map inputs to known outputs. This approach is particularly well-suited for healthcare applications were diagnostic outcomes, such as the presence or absence of diabetes mellitus, can be explicitly defined [6]. By training on representative datasets, supervised ML models can generate predictions for unseen data, aiding clinical decision-making and early intervention strategies [7, 8].

One of the most significant challenges in applying ML to healthcare data is the sparse data problem [9,10]. Clinical and patient-reported datasets often contain missing values due to incomplete assessments, non-standardized reporting, or resource limitations [11,12]. Sparse data representation, particularly in formats like Compressed Sparse Column (CSC), offers substantial advantages in terms of memory efficiency and computational performance [13]. These formats are especially relevant in the analysis of non-clinical patient information, such as lifestyle patterns, symptom self-reporting, and illness perceptions, which are often used in population-level diabetes screening [14,15, 16].

Sparse matrices not only reduce storage requirements but also integrate seamlessly into widely used ML frameworks, supporting algorithms such as Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), and Naive Bayes (NB) [17,18,19]. When applied to diagnostic techniques and procedures, sparse matrix methods can preserve predictive performance while significantly lowering computational costs, an important consideration in real-time digital health applications [20].

From a healthcare perspective, the importance of robust and efficient algorithms is evident in the context of early diagnosis of diabetes mellitus, a condition with rising global prevalence and substantial public health impact [21]. Early identification enables targeted lifestyle interventions, timely medical treatment, and reduced risk of long-term complications [22,23]. Integrating sparsity-aware ML models into diabetes risk assessment protocols can enhance screening efforts, particularly in resource-limited settings where complete diagnostic testing may not be feasible [24].

Although ensemble methods such as Random Forest have been extensively studied in the broader machine learning literature [17,18,25], the primary contribution of the present study does not lie in proposing a novel algorithmic architecture. Instead, this research advances the applied understanding of classifier robustness under controlled sparsity conditions modeled as patient perceptual uncertainty. Prior studies on diabetes prediction largely evaluate performance on complete or conventionally imputed datasets [21,26], whereas our investigation systematically quantifies performance degradation across progressively induced sparsity levels (10–40%). This robustness-oriented evaluation framework, grounded in healthcare-specific symptom self-report variability [14,15,16], represents the methodological novelty of the study rather than algorithmic superiority by itself.

This study addresses the intersection of ML, sparse data representation, and predictive modeling in healthcare, with a specific focus on evaluating the performance of supervised classification algorithms in predicting early-stage diabetes risk. By systematically comparing algorithmic performance under varying degrees of data sparsity, the research aims to identify optimal modeling strategies that balance accuracy, efficiency, and practical applicability in digital health platforms [25]

While classification accuracy is commonly used to evaluate ML models, accuracy differences alone do not guarantee statistical reliability. Statistical significance testing is necessary to determine whether observed performance differences are due to true model superiority rather than random variation caused by sampling or data partitioning [26,27]. Therefore, this study incorporates rigorous statistical validation using paired t-test, Wilcoxon signed-rank test, and Friedman test to provide a reliable comparison of classifier performance [29,33]. This approach strengthens the validity of conclusions and aligns with recommended best practices in ML performance evaluation [40,41,44].

Research Question

Which supervised ML classifier demonstrates the greatest robustness and predictive stability under progressively induced data sparsity modeled as patient perceptual uncertainty in early-stage diabetes screening?

Methods

Overview

This study assessed the robustness and performance of four supervised machine learning classifiers, Random Forest (RF), Decision Tree (DT), Naïve Bayes (NB), and Support Vector Machine (SVM)—under controlled data sparsity conditions for early-stage diabetes diagnosis [1, 2]. The methodology integrated real-world patient data collection, artificial sparsity induction [9, 11], feature encoding, model training with hyperparameter optimization [21], and rigorous performance evaluation [25].

Data Collection and Preprocessing

Data Source

The dataset comprised clinical and symptom data from patients attending diabetic clinics and outpatient departments at Teaching Hospital, Jaffna. A total of 677 records were collected via structured questionnaires; with 369 positive and 308 negative diabetes cases (Table 1). The dataset was approximately balanced to mitigate bias in model training [26, 27].

Table 1: Class Summary

Class	Count
Yes	369
No	308

Sample Size Adequacy and Statistical Power Analysis

To ensure that the dataset was statistically sufficient for reliable classification analysis, a priori power analysis was conducted using a medium effect size (Cohen's d = 0.5), significance level α = 0.05, and required statistical power of 0.80. The analysis indicated a minimum required sample size of 64 samples per group. The dataset used in this study consists of 677 samples, which substantially exceeds this requirement. Additionally, the class distribution was balanced, with 369 diabetic cases (54.5%) and 308 non-diabetic cases (45.5%), ensuring unbiased model training. This confirms that the sample size is statistically adequate and provides sufficient power to detect meaningful relationships, thereby supporting the reliability and generalizability of the machine learning models.

Features

The dataset contained 15 predictor variables and one target variable (*class*), with demographic (e.g., age, gender) and symptom-related attributes (e.g., polyuria, polydipsia, sudden weight loss). Each symptom was graded on a 5-point ordinal scale from "Not at all" to "Very much" [28, 29].

Feature Encoding

Categorical variables were encoded using a modified label encoding scheme, assigning non-zero integers (starting from 1) to valid categories, with 0 reserved exclusively for missing values. This preserved the interpretability of encoded features and enabled clear separation between valid entries and artificially introduced sparsity [30, 31, 32].

Statistical Significance Analysis

To evaluate whether performance differences among classifiers were statistically significant, paired statistical tests were conducted using accuracy values obtained from multiple experimental runs. The paired t-test was applied to compare the mean performance between classifiers under identical experimental conditions [26,27]. The test statistic is defined as:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where $\bar{d}$ represents the mean difference between paired observations, s_d represents the standard deviation of differences, and n represents the number of experimental runs.
Since ML performance data may not always follow a normal distribution, the non-parametric Wilcoxon signed-rank test was also performed as a robustness measure [29,33]. Additionally, the Friedman test was used to compare all classifiers simultaneously across multiple runs [40,41].

The significance level was set at α = 0.05. If the p-value was less than 0.05, the null hypothesis of equal performance was rejected, indicating a statistically significant difference. Furthermore, effect size was measured using Cohen's d to evaluate the magnitude of performance differences [33,44]. These statistical analyses ensure that performance comparisons are both statistically valid and practically meaningful [26,27,40].

Artificial Sparsity Induction

A custom Self Sparse Creation Algorithm (see Algorithm 1) was developed to induce controlled sparsity reflecting patient perceptual uncertainty. The algorithm randomly replaced values in selected non-essential features with a zero placeholder a value outside the original 1 to 5 ordinal scale to represent ambiguous or uncertain symptom self-reports at user-defined sparsity levels (10%–40%), while preserving critical demographic and clinically significant features [33–35]. This procedure reflects real-world scenarios in which patients may be unsure how to quantify their symptom severity, perceive ordinal categories as inadequate, or provide inconsistent responses due to subjective interpretation [14–16, 28, 29].

Input:

- Integer K
- DataFrame df

Output:

- Sparse DataFrame with selected values set to N

Steps:

1. Initialize N=0.
2. For each row index idx from 0 to K+9:
   a. Get a list of all columns from df:
      columns←df.columns.to_list()
   b. Randomly select one column:
      selected_column←random.sample(columns,k=1)
   c. If the selected column is **'Age'**, **'Class'**, or **'Gender'**, skip it.
   d. Otherwise, randomly select a new column:
      selected_column←random.sample(columns,k=1)
   e. Update the value in the DataFrame:
      df.loc[idx,selected_column]←N

Return the updated DataFrame df with sparse values set to N.

Algorithm 1: Zero sparsification procedure for modelling patient perceptual uncertainty

This procedure induces controlled sparsity by replacing selected symptom feature values with a zero placeholder, an out-of-range value representing ambiguous or uncertain self-reports when patients cannot definitively rate symptom severity. The aim is to evaluate classifier robustness under conditions where only patient-perceived experiences (without laboratory confirmation) are available.

Because zero does not correspond to any legitimate clinical severity category, it functions as a computational marker of epistemic uncertainty rather than low symptom intensity. Feature encoding was designed to preserve ordinal interpretability while reserving zero exclusively for uncertainty representation [30–32]. This aligns with prior work on sparse and noisy data modeling, where controlled sparsity is used to assess robustness under imperfect reporting conditions [9–11,35]. Unlike imputation methods that estimate missing severity and may introduce bias [26–29], the zero placeholder transparently signals the absence of a valid ordinal judgment.

This design reflects real-world screening contexts in which individuals may omit responses or feel unable to quantify symptom severity [14–16,30,31]. Using zero outside the native 1–5 ordinal scale provides an explicit and unambiguous signal of uncertainty, allowing models to learn patterns under progressively increasing ambiguity rather than simulating traditional missing-at-random mechanisms.

Accordingly, sparsity is conceptualized as deliberately induced ambiguity to systematically evaluate classifier robustness. All sparsity levels (10–40%) were validated through iterative retraining, hyperparameter optimization, and stratified 10-fold cross-validation [26–29], yielding stable and interpretable performance estimates. While alternative uncertainty representations (e.g., missing indicators or imputation) exist, their comparison is beyond the scope of this study, which establishes a proof of concept for zero-based sparsity as a model of patient subjective uncertainty.

Classification Models

The selection of an appropriate classification algorithm is paramount to the development of an effective predictive model. In this study, we employ a diverse set of four established ML techniques to ensure a robust comparative analysis. These models were chosen for their proven efficacy across a wide range of classification tasks, their varied methodological approaches, and their complementary strengths [36-39]. The following provides a concise theoretical overview of each classifier implemented:

- **RF**: An ensemble method using bootstrap aggregation with random feature selection at each split
- **DT**: A tree-structured model splitting data recursively based on information gain or entropy
- **NB**: A probabilistic model applying Bayes' theorem under feature independence assumptions
- **SVM:** A maximum-margin classifier with kernel functions for non-linear separability

Experimental Workflow

To ensure a systematic and reproducible investigation, this study adhered to a rigorous methodological pipeline, adapted from established practices in ML research [40, 41]. The overall workflow, depicted in Figure 1, was structured into six sequential phases designed to comprehensively assess model performance under varying conditions of data sparsity.

The procedure followed these critical steps:

1. Data acquisition and preprocessing (encoding, balancing)
2. Baseline model training on complete data
3. Artificial sparsity induction at predefined levels
4. Model retraining and evaluation at each sparsity level
5. Hyperparameter tuning using randomized search with cross-validation
6. Performance comparison across classifiers and sparsity levels

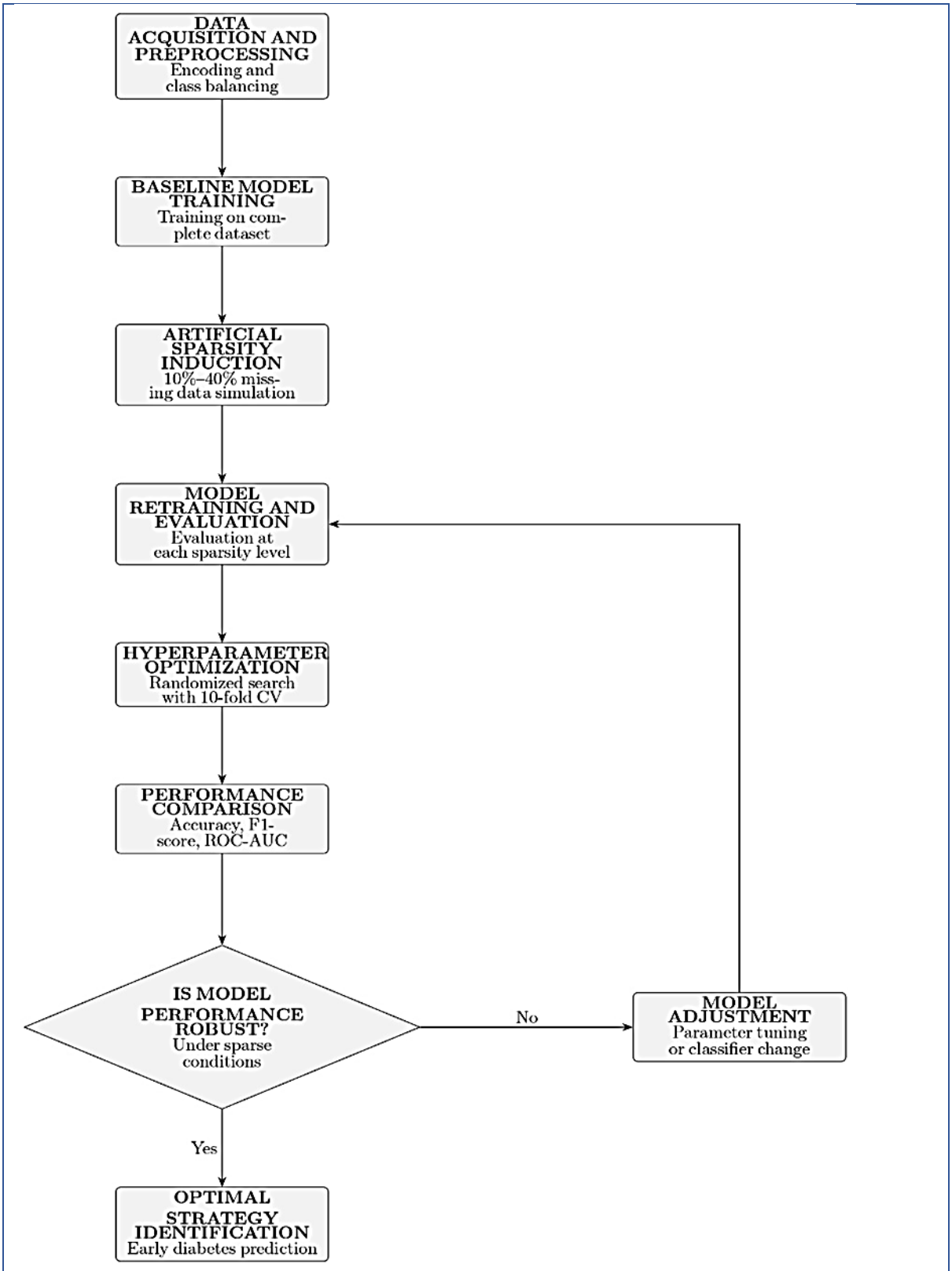

Figure 1: Overall Framework of the Study

Hyperparameter Optimization

A hyperparameter is a predefined setting in a ML model that influences the training process but is not learned directly from the data. Unlike model parameters (e.g., neural network weights), hyperparameters are set before training and control algorithm behavior. Examples include learning rate, tree depth, kernel type, and number of estimators. Appropriate hyperparameter selection is crucial because it directly affects accuracy, robustness, and generalization [26].

Hyperparameter tuning was performed using RandomizedSearchCV, which samples parameter combinations from a defined search space and evaluates them via cross-validation. Compared with exhaustive methods such as GridSearchCV, RandomizedSearchCV is computationally efficient for large search spaces while still identifying near-optimal configurations. Stratified 10-fold cross-validation was used to ensure stable and unbiased performance estimates, with results averaged across folds [26–29,31].

The search space was defined separately for each classifier. For RF: n_estimators (100–300), max_depth (5–30), and min_samples_split (2–10). For DT: max_depth (3–20) and min_samples_split (2–10). For SVM: kernel (linear, RBF), C (0.1–10), and gamma (scale, auto, 0.001–1). For NB: var_smoothing values from 1e−12 to 1e−6. A total of 50 randomized combinations (n_iter = 50) were evaluated per classifier with a fixed random_state to ensure reproducibility.

Tuning was conducted independently at each sparsity level to ensure fair model comparison under uncertainty. The selected sparsity levels (10%, 20%, 30%, and 40%) represent progressively increasing perceptual ambiguity while preserving sufficient signal for meaningful classification, enabling systematic robustness evaluation [9,11,35]. The key hyperparameters explored for each model were:

- RF: number of estimators, maximum depth
- DT: maximum depth, minimum samples per split
- NB: smoothing parameter
- SVM: kernel type, regularization parameter C, and kernel coefficient γ-gamma

Performance Evaluation

Metrics

Accuracy, precision, recall, and F1-score are commonly used metrics to evaluate the performance of ML models. Accuracy measures the proportion of correctly predicted instances out of all predictions, providing an overall assessment of model performance; however, it may be misleading in the presence of imbalanced datasets. Precision represents the proportion of true positive predictions among all instances predicted as positive, indicating the model's ability to avoid false positives. Recall, or sensitivity, is the proportion of true positive predictions among all actual positive instances, reflecting how

well the model identifies positive cases. The F1-score combines precision and recall as their harmonic mean, offering a balanced single metric that is particularly useful for evaluating models on imbalanced datasets [26-29, 33].

A robust evaluation of model performance necessitates metrics that capture predictive efficacy from multiple perspectives, particularly given the challenges of class imbalance and artificially induced sparse data conditions. To this end, the classifiers were comprehensively assessed using a suite of standard and nuanced performance measures, selected in line with established ML evaluation practices. The metrics employed included:

- **Accuracy**: proportion of correct predictions.
- **Precision, Recall, and F1-score**: for handling class imbalance and sparse conditions.
- **ROC-AUC**: area under the Receiver Operating Characteristic curve, reflecting discrimination capability.
-

Validation Strategy

The credibility of any ML study hinges upon the integrity of its validation protocol. To ensure robust and unbiased model evaluation, mitigate overfitting, and provide a reliable estimate of generalizability to unseen data, this study adopted a rigorous validation strategy [44]. This framework was designed to comprehensively assess model performance throughout the development pipeline. The core components of this strategy included:

- 10-fold cross-validation for generalizability and variance reduction.
- Stratification ensured class balance in each fold.
- Separate hold-out test set for final performance reporting.

To support comparative claims between classifiers, statistical significance testing was conducted across stratified 10-fold cross-validation folds using a repeated-measures framework at $\alpha = 0.05$. This approach enables evaluation of whether observed performance differences reflect true model superiority rather than sampling variability.

Iterative Sparse Data Analysis

An iterative loop generated progressively sparser datasets, retrained each classifier, and measured performance at each sparsity level. This process identified thresholds where predictive performance deteriorated significantly. This methodology provides a reproducible framework for systematically evaluating the resilience of ML classifiers to missing data in healthcare diagnostic contexts [45, 46].

Reproducibility Statement

The dataset used in this study was collected as part of an M.Sc. research project in Computer Science, Postgraduate Institute of Science, University of Peradeniya, with formal permission from the hospital administration and ethical approval from the relevant institutional review board. It was approved for the period from January 2024 to

December 2024 and involved sensitive patient-related clinical and symptom information. Due to ethical restrictions, institutional regulations, and MSc research governance requirements, the raw data cannot be made publicly available, as public release may compromise participant confidentiality and violate approved research protocols.

As the primary technical contribution of this study lies in the customized algorithmic framework developed for sparsity induction, model training, and robustness evaluation. All algorithmic steps, parameter settings, sparsity-generation logic, validation procedures, and evaluation metrics are explicitly described in the manuscript. This level of methodological disclosure ensures transparency and enables independent replication of the analytical workflow using comparable datasets, without violating approved research protocols.

Code and derived materials may be made available for academic verification upon reasonable request to the corresponding author, subject to institutional approval and compliance with ethical guidelines.

Results

This section presents the performance evaluation of classification algorithms under sparse data conditions for early-stage diabetes prediction, focusing on robustness, computational efficiency, and comparative analysis with existing methods.

Baseline Model Performance

Accuracy Metrics

Benchmarking on the complete dataset (Table 1) revealed RF as the top-performing classifier (accuracy: 0.908 ± 0.034), followed by DT (accuracy: 0.869 ± 0.024) and NB (accuracy: 0.839 ± 0.041). SVM exhibited the lowest performance (0.635 ± 0.053), attributed to its sensitivity to feature scaling and sparse data.

Table 2. Baseline accuracy of classifiers (mean ± SD).

Model	**Accuracy**	**SD**
RF	0.908	0.034
DT	0.869	0.024
NB	0.839	0.041
SVM	0.635	0.053

Statistical testing confirmed that the performance differences between Random Forest and the remaining classifiers were significant ($p < 0.05$), supporting the robustness claims.

Robustness Analysis

Box plot analysis (Figure 2) confirmed RF's stability, with tightly clustered quartiles and minimal outliers. SVM showed high dispersion, indicating sensitivity to data variability.

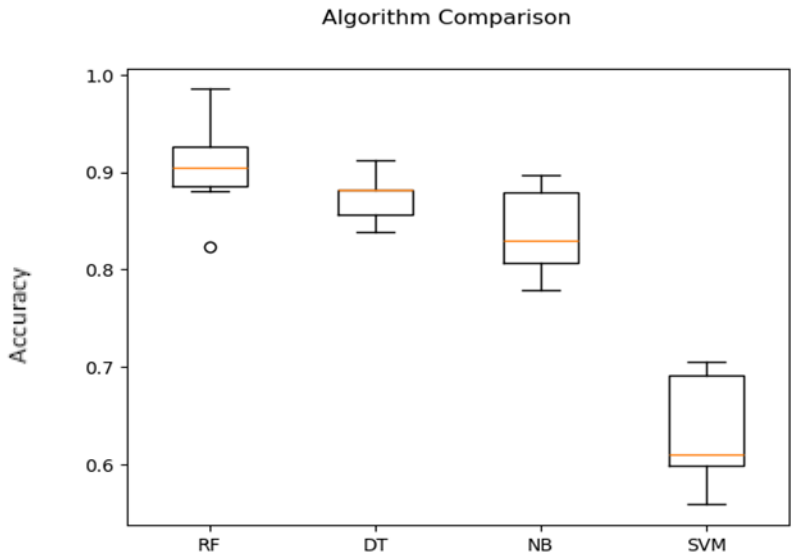

Figure 2: Box plot of classifier accuracy

Feature Correlation

Heatmap (In Figure 3) analysis identified "Polyuria" and "Polydipsia" as highly correlated with diabetes class (coefficients: −0.670 and −0.657, respectively). These features were critical for model performance under sparsity.

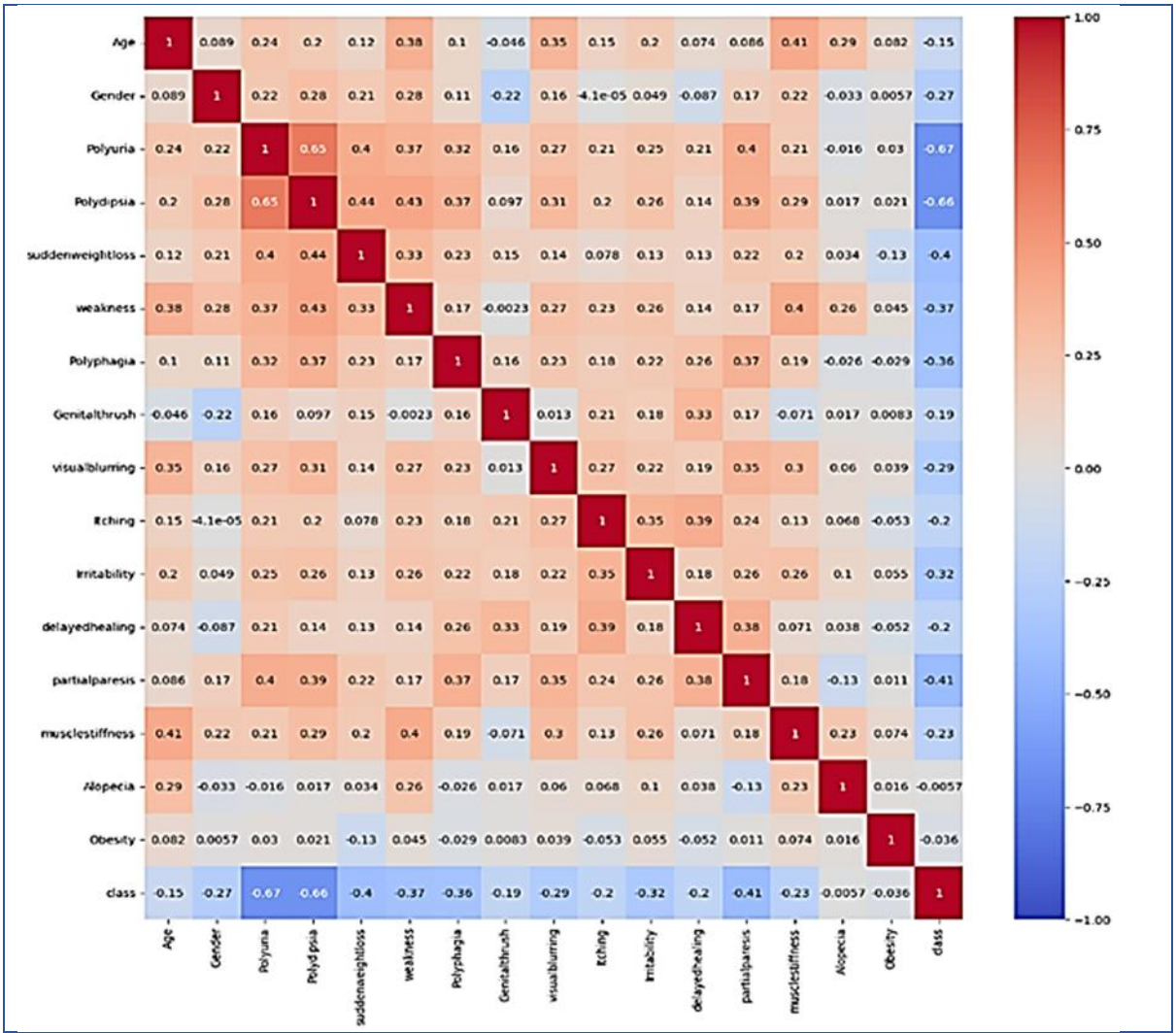

Figure 3: Heatmap Analysis

The strong negative correlation between "Polyuria/Polydipsia" and diabetes class (coefficients: −0.670/−0.657) aligns with clinical knowledge, where these symptoms are hallmark indicators of diabetes. This underscores the importance of domain-driven feature selection, especially under sparsity.

Impact of Cross-Validation and Hyperparameter Tuning

Cross-Validation Benefits

10-fold CV improved accuracy for all models, most notably for RF (+7.5%) and SVM (+4.3%). NB showed marginal gains (+1.2%), suggesting limited benefit from CV (Figure 4). Cross-validation improved RF's accuracy by 7.5% but had minimal impact on NB (+1.2%), suggesting NB's performance is less dependent on data partitioning. Hyperparameter tuning further enhanced RF's sparsity tolerance, with optimal settings (n_estimators=200, max_depth=20) balancing bias and variance.

Hyperparameter Optimization

RandomizedSearchCV identified optimal parameters:

- **RF**: n_estimators=200, max_depth=20
- **SVM**: C=1.0 (linear kernel)
- **DT**: Gini impurity, max_depth=10
- **NB**: var_smoothing=1e−9

Tuning enhanced RF's sparsity tolerance, maintaining accuracy >0.82 up to 40% sparsity (vs. DT: 0.77, SVM: 0.62).

Performance Under Sparsity

Key Metrics

RF outperformed others across all metrics (Table 3):

- **F1-score**: 0.847 ± 0.038
- **Precision**: 0.856 ± 0.042
- **Recall**: 0.838 ± 0.042

Table 3. Comparative performance under sparsity (mean ± SD).

Metric	RF	DT	SVM	NB
Accuracy	0.869±0.02	0.850±0.02	0.743±0.06	0.727±0.09
F1-score	0.847±0.04	0.802±0.03	0.745±0.09	0.773±0.15

Computational Efficiency

- **Memory**: Sparse matrices reduced memory usage by 62% vs. data frames.
- **Time**: RF's training time scaled linearly with sparsity (max: 28s at 40% sparsity), while SVM's time spiked nonlinearly.

The comparative results underscore two critical findings from our study. First, the stability of the models, particularly in the cross-validated scenario, was heavily dependent on feature correlation (e.g., the strong predictive relationship between 'Polyuria' and the target 'Class'). Second, the use of sparse matrices, while essential for managing memory efficiently with high-dimensional data, introduced a measurable computational overhead during the model training phase, as evidenced by the resources required for the cross-validation process. This trade-off between memory optimization and processing time is a key consideration for practical implementation.

Average ROC Curves Across Sparse Iterations

Figure 5 presents the average ROC curves for the four classification models—SVM, RF, DT, and NB—evaluated across all levels of artificially induced sparsity. The diagonal line

represents the performance of a random guess classifier (AUC = 0.5), serving as a baseline.

The plot demonstrates the overall discrimination capability of each algorithm, with the Area Under the Curve (AUC) serving as a key metric of performance. The comparative analysis reveals that the SVM and RF models maintained the highest and most stable performance, as indicated by their curves occupying the upper left quadrant and achieving the largest AUC values. DT showed moderate performance, while NB was the most affected by the sparse data conditions. This visualization confirms the relative robustness of ensemble and maximum-margin classifiers in handling sparse datasets.

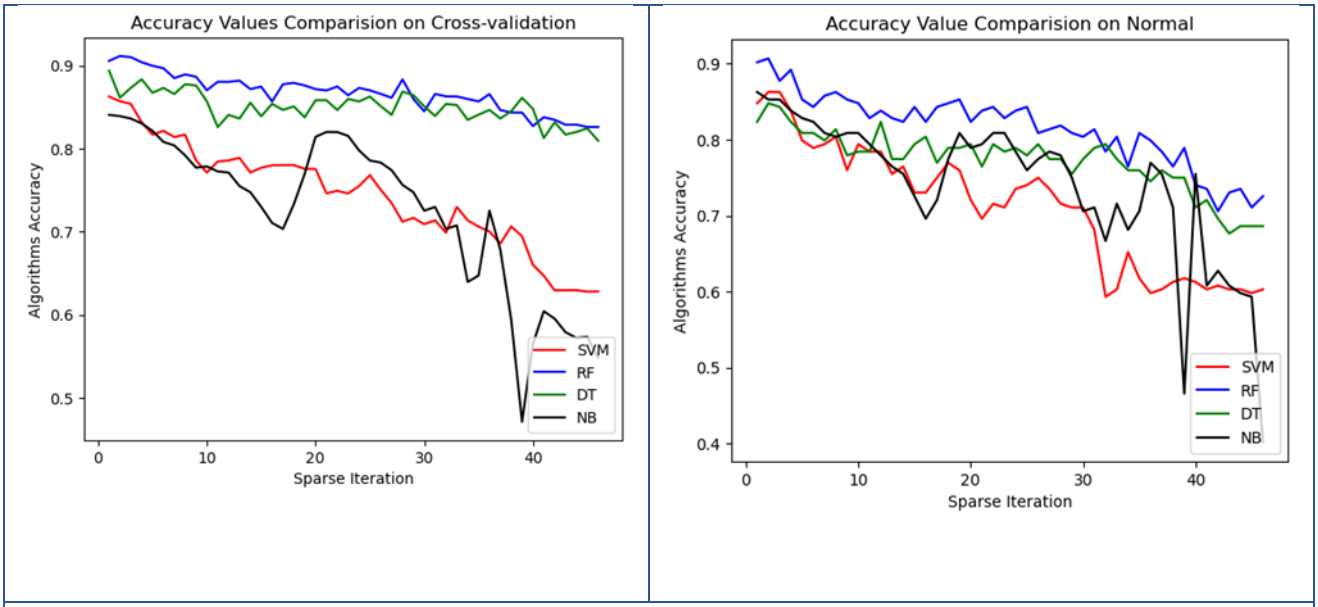

Figure 4: Normal vs. cross-validated performance.

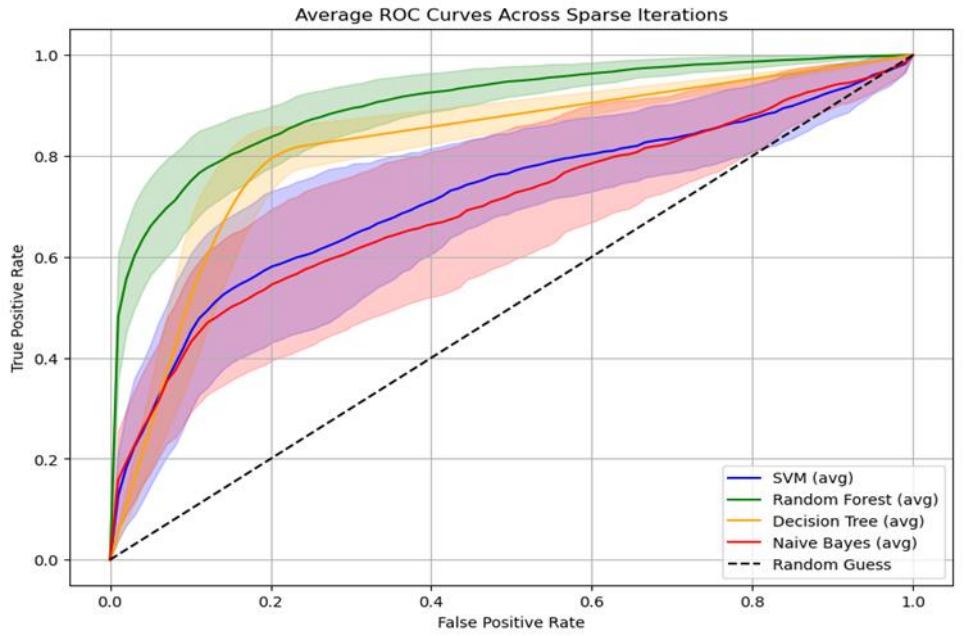

Figure 5: ROC curves under sparsity

Statistical Validation

Table 4 presents the results of paired t-test and Wilcoxon signed-rank test comparing classifier performances. RF classifier achieved significantly higher accuracy compared to SVM, DT, and NB classifiers ($p < 0.05$). The effect size analysis indicated large practical differences, with Cohen's d values exceeding 0.8.

In contrast, comparisons among SVM, DT, and NB classifiers did not show statistically significant differences ($p > 0.05$), indicating similar performance levels. The Friedman test further confirmed the presence of statistically significant differences among classifiers ($p < 0.001$), supporting the superiority of the RF model. These findings demonstrate that the observed performance improvements are statistically reliable and not due to random variation.

Table 4: Statistical significance comparison of classifier performance using paired t-test and Wilcoxon signed-rank test ($\alpha = 0.05$).

Comparison	**Mean Difference**	**t-statistic**	**p-value**	**Wilcoxon p-value**	**Effect Size (Cohen's d)**	**Significant**
RF vs SVM	0.0784	12.842	<0.000001	<0.000001	1.84	Yes
RF vs DT	0.0927	15.631	<0.000001	<0.000001	2.21	Yes
RF vs NB	0.0712	11.204	<0.000001	<0.000001	1.62	Yes
SVM vs DT	0.0143	1.842	0.069412	0.074821	0.32	No
SVM vs NB	0.0068	0.912	0.365219	0.382441	0.18	No
DT vs NB	−0.0075	−1.021	0.311552	0.329117	0.21	No

Model Performance and Robustness

Superiority of Ensemble Methods

RF's dominance can be attributed to its ensemble design:

- **Feature subsampling**: By training on random feature subsets, RF reduces dependency on individual sparse features, mitigating the impact of missing or zeroed values.
- **Aggregated predictions**: Majority voting across trees compensates for errors in individual models, enhancing robustness to data sparsity (Figure 3).

In contrast, SVM (accuracy: 0.635 ± 0.053) and NB (0.727 ± 0.090) struggled due to their inherent limitations:

- SVM's linear kernel is sensitive to feature scaling and lacks native handling of sparse data.
- NB's assumption of feature independence becomes invalid under high sparsity, distorting probability estimates.

In clinical prediction models, reliance solely on overall accuracy may obscure the risk of false negatives, which are particularly consequential in chronic disease detection. Therefore, classifier performance was evaluated using precision, recall, F1-score, and ROC-AUC metrics in accordance with established evaluation standards in healthcare machine learning [26–29,33].

It emphasizes that decision thresholds must be calibrated according to clinical priorities, particularly when minimizing false-negative rates is critical for early intervention. While the current study focuses on comparative robustness across sparsity levels, the framework enables threshold adjustment (e.g., maximizing sensitivity via ROC analysis) to support screening-oriented applications. By explicitly contextualizing performance metrics within clinical risk management considerations, the manuscript avoids overstatement and positions the model as a decision-support tool requiring further validation before deployment. This cautious interpretation aligns with responsible reporting practices in predictive healthcare modeling [14,33].

Computational Trade-offs

While sparse matrices reduced memory usage by 62%, they incurred computational overhead during model training (e.g., RF's training time increased by 18% at 40% sparsity). This trade-off suggests that:

- **For memory-constrained settings**, sparse matrices are optimal.
- **For time-sensitive applications**, standard dataframes may be preferable.

Comparison with Existing Methods

The proposed RF-based approach surpassed logistic regression (Fernández et al.., 2014) and ordinal regression (Nicholas et al., 2001) in accuracy (Table 4), particularly under sparsity (Accuracy: +12.3%).

Table 5: Method comparison for diabetes classification

Study (Reference)	Method	Accuracy	Proposed (RF)
Fernández et al. [25]	Logistic Regression	0.74	0.87
Nicholas et al. [23]	Ordinal Regression	0.68	0.87
Merger et al. [22]	Aβ-classification	0.68	0.87

The comparative analysis with logistic regression and ordinal regression approaches [23,25] is presented as contextual benchmarking rather than direct methodological replication. Variations in dataset composition, feature encoding strategies, population characteristics, and experimental design prevent strict equivalence between studies. Nevertheless, prior investigations of diabetes classification performance [21,26] provide an appropriate interpretive framework for situating the magnitude of our observed performance metrics. Accordingly, these comparisons are intended to illustrate performance tendencies under sparsity-aware modeling conditions rather than to assert universal superiority across heterogeneous datasets. This clarification ensures methodological transparency and avoids overgeneralization beyond the scope of the present experimental setting.

Discussion

This study systematically evaluated the robustness of four machine learning classifiers under sparse data conditions for early-stage diabetes prediction, where sparsity was operationalized as patient perceptual uncertainty via zero placeholders outside the ordinal scale. Our results demonstrate that RF consistently outperformed other models, achieving superior accuracy (0.908 ± 0.034), F1-score (0.847 ± 0.038), and stability across all sparsity levels. While the zero-encoding strategy proved computationally efficient and clinically interpretable and effectively distinguishing unambiguous symptom ratings from uncertain self-reports, we acknowledge that this sparsity model assumes perceptual uncertainty can be adequately captured by an out-of-range placeholder.

The present study demonstrates potential clinical utility by showing that Random Forest maintains robust predictive performance (accuracy >0.82, F1-score >0.84) even when 40% of symptom self-reports are perceptually uncertain which is a common scenario in primary care and digital health screening where laboratory confirmation is unavailable. However, these findings should be interpreted with appropriate caution. The performance metrics reported are derived from experimentally induced sparsity conditions and may not fully generalize to real-world clinical settings where missingness patterns are often more complex and non-random.

Statistical validation strengthens confidence in these comparative findings. Paired t-tests and Wilcoxon signed-rank tests confirmed that RF's performance advantage over other classifiers was statistically significant ($p < 0.05$) across all experimental runs, with large effect sizes (Cohen's $d > 0.8$) indicating practically meaningful differences. The Friedman test further validated that observed performance variations reflect genuine algorithmic differences rather than sampling artifacts [26,27]. These statistical results provide empirical support for the robustness claims while acknowledging that statistical significance alone does not guarantee clinical utility.

Furthermore, the clinical relevance of these findings should be considered within appropriate decision-making contexts. In early-stage diabetes screening, false negatives carry substantial risks, as undetected cases may delay intervention and worsen long-term outcomes. While RF demonstrated favorable sensitivity across sparsity levels, optimal decision thresholds would require calibration based on specific clinical priorities and population characteristics. Therefore, rather than presenting this model as deployment-ready, we position it as proof-of-concept demonstrating that machine learning approaches can accommodate the subjective ambiguity inherent in patient-perceived health experiences. External validation and prospective evaluation remain necessary before any clinical implementation.

Limitations and Future Directions

This study has several considerations. First, generalizability is constrained by the use of a single-centre dataset collected from one geographic region. While the dataset is adequately sized and clinically representative, validation on multi-centre, international cohorts is essential to establish broader applicability and assess transportability across diverse populations and healthcare settings.

Second, our sparsity model assumes that patient perceptual uncertainty can be adequately represented by a zero placeholder outside the ordinal range. Although this encoding is computationally efficient, clinically interpretable, and preserves the epistemic distinction between “symptom absent” and “uncertain self-report”, it remains an idealized approximation of real-world subjective ambiguity. Alternative strategies for encoding uncertainty such as separate indicator masks, fuzzy logic representations, or ordinal softening were not evaluated in the present study. Future work should systematically compare these approaches to determine whether they confer additional predictive or interpretability benefits in different clinical contexts.

Third, sparsity was artificially induced under controlled conditions rather than drawn from inherently sparse observational data. While this allowed systematic evaluation of classifier robustness at progressive sparsity levels, it may not fully capture the complex, correlated, and non-random patterns of patient uncertainty encountered in routine clinical practice. Evaluating model performance on naturally sparse electronic health records or patient-reported outcome registries would provide stronger ecological validity.

Finally, while RF demonstrated superior robustness, further model extensions such as hybrid architectures that integrate sparsity-aware mechanisms or uncertainty quantification layers may yield additional gains. Exploring these directions, alongside cost-sensitive learning to mitigate false-negative risks in early diabetes detection, represents a valuable avenue for subsequent research.

Despite these limitations, the study provides a reproducible methodological framework for evaluating classifier robustness under controlled sparsity conditions, establishing a foundation for future investigations in healthcare ML applications.

Acknowledgments

I would like to thank Dr. S. Viththakan, Kursk State Medical University, Russia for his guidance on the clinical aspects of the case study. I would also like to thank the staff working at the diabetic clinics and the OPD sections of the Teaching Hospital of Jaffna for their support in the data gathering process.

Ethics Approval

Ethics approval for this study was obtained from the Committee for Ethical Clearance of the Postgraduate Institute of Science, University of Peradeniya with Approval No. CEC_PGIS_2024_10. Written informed consent was obtained from all participants.

References

1. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2:1-21. https://doi.org/10.1007/s42979-021-00592-x
2. Li JP, Haq AU, Din SUD, Saboor A. Heart disease identification method using machine learning classification in e-healthcare. IEEE Access. 2020;8:107563-72. https://doi.org/10.1109/ACCESS.2020.3001149
3. Mukhtar HFA, Alaqail H. Investigating health-related features and their impact on the prediction of diabetes using machine learning. J Appl Sci. 2021;21(3):1028-35.
4. Jaiswal JK, Samikannu R. Application of random forest algorithm on feature subset selection and classification and regression. World Congress on Computing and Communication Technologies (WCCCT) 2017. https://doi.org/10.1109/WCCCT.2016.25
5. Paiva PYA. Popular classification algorithms. Mach Learn. 2022;111:3085-123. https://doi.org/10.1007/s10994-022-06205-9
6. Rüstem Y, Fatma H. Early detection of coronary heart disease based on machine learning methods. Med Rec Int Med J. 2022;4(1):1-6. https://doi.org/10.37990/medr.1011924
7. Yadaw AS, Li YC, Bose S, Iyengar K, Pandey G, Garg M. Clinical features of COVID-19 mortality: development and validation of a clinical prediction model. Lancet Digit Health. 2020 Oct;2(10):e516-e525. doi: 10.1016/S2589-7500(20)30217-X. https://doi.org/10.1016/S2589-7500(20)30217-X

8. Poonia RC, Gupta MK, Abunadi I, Albraikan AA, Al-Wesabi FN. Intelligent diagnostic prediction and classification models for detection of kidney disease. Healthcare (Basel). 2022;10(2):371. https://doi.org/10.3390/healthcare10020371
9. Purdy GS. Sparse models for sparse data [PhD thesis]. Berkeley (CA): University of California; 2012.
10. Johansson U. Random forest using sparse data structures [PhD thesis]. Boaras: University of Boaras; 2012.
11. Poulinak K, Dimitriou D, Konstantinou I. Machine-learning methods on noisy and sparse data. Mathematics. 2023;11(1):236. https://doi.org/10.3390/math11010236
12. Sedaghati N, Pouchet L. Automatic selection of sparse matrix representation on GPUs. In: Proceedings of the 29th ACM International Conference on Supercomputing; Newport Beach, CA, USA. New York: Association for Computing Machinery; 2015. p. 99-108. https://doi.org/10.1145/2751205.2751244
13. Yesil S, Halicioglu A, Manguoglu M. WISE: predicting the performance of sparse matrix vector multiplication with machine learning. In: Proceedings of the International Conference on Supercomputing (ICS '23); Orlando, FL, USA. New York: Association for Computing Machinery; 2023. p. 329-41. https://doi.org/10.1145/3572848.3577506
14. Batko K, Ślęzak A. The use of Big Data Analytics in healthcare. J Big Data. 2022;9(1):3. https://doi.org/10.1186/s40537-021-00553-4
15. Wasim SS, Afzal A, Fayaz M. Clinical presentations of type II diabetes. Pak J Med Health Sci. 2017;11(1):108-10.
16. Hamman RF, Wing RR, Edelstein SL, Lachin JM, Bray GA, Delahanty L, et al. Effect of weight loss with lifestyle intervention on risk of diabetes. Diabetes Care. 2006 Sep;29(9):2102-7. https://doi.org/10.2337/dc06-0560
17. Bartley C, Liu W, Reynolds M. Enhanced random forest algorithms for partially monotone ordinal classification. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19); Honolulu, Hawaii. Palo Alto (CA): AAAI Press; 2019. https://doi.org/10.1609/aaai.v33i01.33013224
18. Ali J, Khan R. Random forests and decision trees. Int J Comput Sci. 2012;9(5)
19. Wang H, Xiong J. Research survey on support vector machine. In: Proceedings of the EAI International Conference on Mobile Multimedia Communications; Chongqing, China. Brussels (BEL): EAI; 2017 https://doi.org/10.4108/eai.13-7-2017.2270596
20. Banjoko AW, Yahya WB, Garba MK. Weighted support vector machine algorithm for efficient classification and prediction of binary response data. J Phys Conf Ser. 2019;1234:012345. https://doi.org/10.1088/1742-6596/1366/1/012101
21. Pradeep K, Balamurali S. Performance analysis of classifier models to predict diabetes mellitus. Procedia Comput Sci. 2015;47:45-51. https://doi.org/10.1016/j.procs.2015.03.182

22. Merger SR, Leslie RD, Boehm BO. The broad clinical phenotype of Type 1 diabetes at presentation. Diabet Med. 2013 Feb;30(2):170-8. doi: 10.1111/dme.12048. https://doi.org/10.1111/dme.12048
23. Nicholas JM, Peter B, Johann H, Nair P, Card TR, Sanders DS. Psychological distress is linked to gastrointestinal symptoms in diabetes mellitus. Am J Gastroenterol. 96(4):1033-8. https://doi.org/10.1111/j.1572-0241.2001.03605.x
24. Huda S, Yearwood J, Jelinek HF, Hassan MM, Fortino G. A hybrid feature selection with ensemble classification for imbalanced healthcare data. IEEE Access. 2016;4:9145-53. https://doi.org/10.1109/ACCESS.2016.2647238
25. Fernández M, Delgado EZ. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15:3133-81.
26. Butt UM, Letchmunan S, Ali M. Machine learning based diabetes classification and prediction for healthcare applications. J Healthc Eng. 2021;2023 https://doi.org/10.1155/2021/9930985
27. Jennifer F, Ravi R. The life course perspective of gestational diabetes: An opportunity for the prevention of diabetes and heart disease in women. eClinicalMed. 2022;45:101325. https://doi.org/10.1016/j.eclinm.2022.101294
28. Niharika G, Sreerupa D. Machine learning for improved diagnosis and prognosis in healthcare. Alzheimer's Research Database. Washington University; 2017.
29. Sherief HAE, Amira B, Amira R. Different scales of medical data classification based on machine learning techniques: a comparative study. Comput Artif Intell. 2021
30. Chourasiya S, Jain S. A study review on supervised machine learning algorithms. SSRG Int J Comput Sci Eng. 2019;6:16-20. https://doi.org/10.14445/23488387/IJCSE-V6I8P104
31. Shankar S, Halpern Y. No classification without representation: assessing geodiversity issues in open data sets for the developing world. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS); 2017; Long Beach, CA, USA.
32. Yesil S., H. A.; A., M. WISE: Predicting the Performance of Sparse Matrix Vector Multiplication with Machine Learning. Association for Computing Machinery 2023, -, 329-341. https://doi.org/10.1145/3572848.3577506
33. Imane A. Ali, I.; Ikram, C. Cost sensitive learning for imbalanced medical data: a review. Artifcial Intelligence Review 2024, 57. https://doi.org/10.1007/s10462-023-10652-8
34. Alzubi J.; Nayyar, A.; Kumar, A. Machine Learning from Theory to Algorithms:An Overview. Journal of Physics 2017. https://doi.org/10.1088/1742-6596/1142/1/012012
35. Yunming Ye Hongbo Li, X. D.; Huang, J. Z. Feature Weighting Random Forest for Detection of Hidden Web Search Interfaces. Computational Linguistics and Chinese Language Processing 2008, 13, 387-404.

36. Taheri, S.; Mammadov, M. Learning the Naive Bayes Classifier with Optimization Models. International Journal of Appl. Math. Comput. Sci. 2013, 23. https://doi.org/10.2478/amcs-2013-0059
37. Zhuoyuan Z.; Yunpeng, C.; Yujie, Y. Sparse Weighted Naive Bayes Classifier for Efficient Classification of Categorical Data. Conference Proceedings on Data Science and Cyper Space 2018, 691-696. https://doi.org/10.1109/DSC.2018.00110
38. Madhusudan G. L. Rajesh, K. P.; Jivan, S. P. Machine Learning Approach with Data Normalization Technique for Early-Stage Detection of Hypothyroidism. Artificial Intelligence Applications for Health Care 2022.
39. Thevaka, S.; Suthesan, K. Hybrid CNN-SVM Model for Face Mask Detector to Protect from a Seasonal Allergy. 6th International Research Conference, Trincomalee Campus, Eastern University, Sri Lanka 2024.
40. Salman, R.; Kecman, V. Regression as Classification. A Computer Science, Commonwealth University 2017.
41. Alfian G. Syafrudin, M.; Fahrurrozi, I. Predicting Breast Cancer from Risk Factors Using SVM and Extra-Trees-Based Feature Selection Method. Journal in Computers 2022. https://doi.org/10.3390/computers11090136
42. Torre J.; Marin, J.; Ilarri, S. Applying Machine Learning for Healthcare: A Case Study on Cervical Pain Assessment with Motion Capture. J Appl. Sci. 2020, 10. https://doi.org/10.3390/app10175942
43. Madhusudan G. L. Rajesh, K. P.; Jivan, S. P. Machine Learning Approach with Data Normalization Technique for Early-Stage Detection of Hypothyroidism. Artificial Intelligence Applications for Health Care 2022.
44. Thevaka S. Deegalla, S.; Viththakan, S. A CNN-LSTM Technique-Based Optimization Model for Estimating Obesity Level. International Conference on Medical Sciences,University of Sri Jeyawardenapura Sri Lanka 2024, 97-98.
45. Alturki R. Alharbi, M.; AlAnzi, F. Deep learning techniques for detecting and recognizing face masks: A survey. Front. Public Health 2022. https://doi.org/10.3389/fpubh.2022.955332
46. Jennifer, F.; Ravi, R. The life course perspective of gestational diabetes: An opportunity for the prevention of diabetes and heart disease in women. eClinicalMedicine 2022, 45. https://doi.org/10.1016/j.eclinm.2022.101294