SERBIAN JOURNAL OF ELECTRICAL ENGINEERING
Vol. 22, No. 1, February 2025, 93-112
UDC: 004.85:616.379-008.64 **DOI: https://doi.org/10.2298/SJEE2501093K**
Original scientific paper

Machine Learning for Early Diabetes Screening: A Comparative Study of Algorithmic Approaches

Adem Korkmaz[1], Selma Bulut[2]

Abstract: Diabetes mellitus, a chronic metabolic disorder, poses a significant global health challenge. Early screening and risk assessment are crucial for effective management and prevention. This study evaluates the performance of various machine learning models – Artificial Neural Networks (ANNs), Random Forest (RF), k-nearest Neighbors (k-NN), and Support Vector Machine (SVM) – in screening diabetes risk using a dataset based on patient-reported symptoms such as age, gender, polyuria, polydipsia, and sudden weight loss. The dataset, comprising self-reported data from 520 individuals, highlights the potential association of specific symptoms and demographics with diabetes risk. Rigorous analysis demonstrates the superior performance of the RF model in terms of accuracy and F1 Score. Feature importance analysis further emphasizes the critical role of patient-reported symptoms in assessing predisposition to diabetes. The findings suggest that with its robust predictive capability, RF is particularly suitable for early screening, offering valuable insights into symptom-based diabetes risk assessment. This research advances non-invasive, symptom-based screening tools, paving the way for early interventions and tailored prevention strategies.

Keywords: Classification Algorithms, Diabetes Detection, Machine Learning, Artificial Intelligence.

1 Introduction

Diabetes mellitus, characterized by chronic hyperglycemia, is one of the most pressing global public health challenges of the 21st century. According to the 2021 International Diabetes Federation (IDF) Diabetes Atlas, approximately 537 million adults aged 20 to 79 live with diabetes. This figure is projected to rise to 643 million by 2030 and 783 million by 2045, highlighting the urgent need for

[1]Bandırma Onyedi Eylül University, Gönen Vocational School, The Department of Computer Technology, Turkey, ademkorkmaz@bandirma.edu.tr, https://orcid.org/0000-0002-7530-7715
[2]Kırklareli University, Vocational School of Technical Sciences, The Department of Computer Technology; Turkey, selma.bulut@klu.edu.tr, https://orcid.org/0000-0002-6559-7704

Colour versions of the one or more of the figures in this paper are available online at https://sjee.ftn.kg.ac.rs

comprehensive and effective strategies to address the growing impact of the disease. The burden of diabetes is particularly severe in low- and middle-income countries, which account for three-quarters of global cases. This disparity underscores the complex relationship between diabetes prevalence and socio-economic factors, further complicating efforts to combat disease globally [1].

Diabetes mellitus is a metabolic disorder affecting over 500 million individuals worldwide and is primarily characterized by high blood glucose levels resulting from impaired insulin production or the body's inability to utilize insulin effectively [1, 2]. Insulin, a hormone produced by pancreatic beta cells, is essential for regulating blood glucose levels by promoting glucose uptake into cells [3]. The disorder manifests in several forms, each with distinct etiologies and clinical features:

Type 1 Diabetes: Caused by the autoimmune destruction of pancreatic beta cells, resulting in an absolute insulin deficiency. While Type 1 diabetes is commonly diagnosed in younger individuals, it can manifest at any age.

Type 2 Diabetes: The more prevalent form, Type 2 diabetes, is primarily driven by insulin resistance coupled with inadequate compensatory insulin secretion. It is often associated with lifestyle factors such as a sedentary lifestyle, poor dietary habits, and obesity [4, 5].

Gestational Diabetes: Occurs during pregnancy and generally resolves after childbirth, but it significantly increases the risk of developing Type 2 diabetes later in life.

Clinically, diabetes manifests through a range of symptoms, including polyuria (increased urination), polydipsia (increased thirst), polyphagia (increased hunger), unexplained weight changes, fatigue, delayed wound healing, and visual disturbances [6]. However, the early stages of the disease often remain undiagnosed, particularly in resource-limited settings where access to advanced diagnostic tools is constrained. This high rate of undiagnosed cases emphasizes the critical need for alternative, non-invasive, and scalable approaches to risk screening, particularly for vulnerable populations.

Recent advancements in machine learning (ML) offer promising opportunities for addressing these challenges. In a study by Katiyar et al. [7], ML and deep learning methods demonstrated significant potential in facilitating the detection and classification of diabetes. These computational approaches leverage large datasets to identify complex patterns and relationships that traditional diagnostic methods may overlook. Building on this foundation, the current study evaluates the efficacy of ML models for non-invasive, symptom-based screening of diabetes risk, utilizing a dataset derived from patient-reported symptoms, including age, gender, polyuria, polydipsia, and sudden weight loss.

Comparison of ML Models: This research compares the performance of four ML algorithms-Artificial Neural Networks (ANNs), Random Forest (RF), k-Nearest Neighbors (k-NN), and Support Vector Machine (SVM)-to determine their effectiveness in symptom-based risk screening. These models are selected for their proven capacity to handle classification tasks and to identify patterns within complex datasets.

Feature Importance Analysis: A detailed feature importance analysis assesses the relative significance of patient-reported symptoms in predicting diabetes risk. This analysis not only identifies key attributes but also guides the development of targeted screening strategies.

By aligning machine learning capabilities with symptom-based data, this study aims to advance non-invasive and scalable tools for diabetes risk assessment. The findings are expected to contribute to the integration of computational models into broader healthcare strategies, ultimately improving early interventions and outcomes for populations disproportionately affected by the disease.

2 Related Work

In the advancing field of machine learning for diabetes detection, diverse studies have significantly contributed to understanding algorithmic efficacies. Komi et al. [8] delved into early diabetes prediction using a suite of data mining techniques, such as ANN, ELM, GMM, SVM, and logistic regression, with their findings elevating ANN as the most effective. Addressing the needs of rural Indian demographics, Ramanujam et al. [9] innovatively developed a multilingual decision support system that amalgamates predictive models with clinical decision support, facilitating both self-assessment and assisted evaluations.

Furthering this domain, Khaleel and Al-Bakry [10] introduced a model grounded in the precision of powerful machine learning algorithms, employing measures like precision, recall, and F1-score and harnessing the Pima Indian Diabetes Dataset (PIDD) to yield significant predictive results with Logistic Regression, Naïve Bayes, and KNN. Tripathi and Kumar [11] investigated early-stage diabetes prediction using algorithms including LDA, KNN, SVM, and RF, with their analysis of the PIDD from the UCI machine learning repository demonstrating Random Forest's superiority in accuracy at 87.66%. Similarly, Kaur and Kumari [12] utilized the R data manipulation tool to create and analyze diverse prediction models on the Pima Indian diabetes dataset, finding the SVM-linear model to outshine others in accuracy and precision.

Xue et al. in [13] employed supervised learning algorithms like SVM, Naive Bayes, and LightGBM on a dataset of 520 diabetic and potential diabetic patients, concluding that SVM outperformed others with a remarkable accuracy rate of

96.54%. Sisodia and Sisodia [14] focused on early-stage diabetes detection using Decision Tree, SVM, and Naive Bayes on the PIDD, with Naive Bayes emerging as the most accurate at 76.30%. Sarwar et al. [15] provided a comparative analysis of various algorithms, including LR, KNN, SVM, RF, and Decision Tree, for diabetes prediction, highlighting the high accuracy of SVM and KNN at 77%. Lastly, Rawat et al. [16] explored multiple Machine Learning Algorithms for diabetes prediction, with their comparative analysis declaring the neural network as the most effective, boasting a 98% accuracy rate.

Collectively, these studies underscore the critical role of machine learning in the early detection of diabetes, revealing how different algorithms can be optimized based on specific data characteristics and desired outcomes, thus paving the way for more effective and tailored approaches in medical diagnostics.

3 Material and Method

In this study, medical data from 520 patients, meticulously collected from the Sylhet Diabetes Hospital in Bangladesh, were utilized. The dataset encompasses responses to a questionnaire with 17 attributes, including demographic details, symptoms, and diabetes-associated risk factors [17]. The data underwent standardization to balance the effects of feature scale differences and were partitioned into an 80% training subset and a 20% testing subset. Various classification models such as Random Forest, k-Nearest Neighbors, Support Vector Machines, and Artificial Neural Networks were employed. These models effectively separate data classes, especially in non-linear contexts, and demonstrate the ability to discern complex patterns across different datasets.

3.1 Dataset

The previously mentioned 17 attributes are shown in **Table 1**. These attributes encapsulate a blend of demographic details, symptoms, and potential risk factors associated with diabetes. The range of attributes includes but is not limited to, age, gender, and specific diabetes-related symptoms such as Polyuria, Polydipsia, and sudden weight loss.

A key classification attribute within this dataset is “Class,” a binary indicator signifying whether a patient has tested positive (1) or negative (0) for diabetes. The dataset consists of 320 positive and 200 negative cases, offering a balanced perspective for analysis. The data collection process adhered to stringent standards, ensuring the reliability and integrity of the data. All entries are non-null and have been validated by certified medical professionals, underscoring the dataset’s accuracy and applicability.

The primary objective of employing this dataset is to derive insights into the symptomatic patterns and correlations inherent in diabetic conditions, with an overarching goal of developing predictive models for diabetes diagnosis [17].

Table 1
Diabetes Dataset Information.

#	Attribute	Data Type	Description
1	Age	int64	Age of the patient
2	Gender	int64	The gender of the patient
3	Polyuria	int64	Presence of excessive urination
4	Polydipsia	int64	Presence of excessive thirst
5	Sudden Weight Loss	int64	Presence of sudden and unexplained weight loss
6	Weakness	int64	Feeling of weakness
7	Polyphagia	int64	Presence of excessive hunger
8	Genital Thrush	int64	Presence of a yeast infection in the genital area
9	Visual Blurring	int64	Presence of blurred vision
10	Itching	int64	Presence of itching
11	Irritability	int64	Presence of irritability
12	Delayed Healing	int64	Presence of slow wound healing
13	Partial Paresis	int64	Presence of partial paralysis
14	Muscle Stiffness	int64	Presence of muscle stiffness
15	Alopecia	int64	Presence of hair loss
16	Obesity	int64	Presence of obesity
17	Class	object	Positive (1) or Negative (0) for diabetes

3.2 Data preprocessing and models

The dataset utilized in this research underwent meticulous preprocessing, normalization, and partitioning steps before implementing classification models. Initially, data normalization was carried out using the X = sc.fit_transform(X) command to standardize the feature variables, which is critical to prevent any individual feature from disproportionately influencing the model due to scale differences. Following normalization, the dataset was strategically divided into an 80% training subset and a 20% testing subset, a fundamental step to effectively evaluate the models' performance and generalization capability on unseen data.

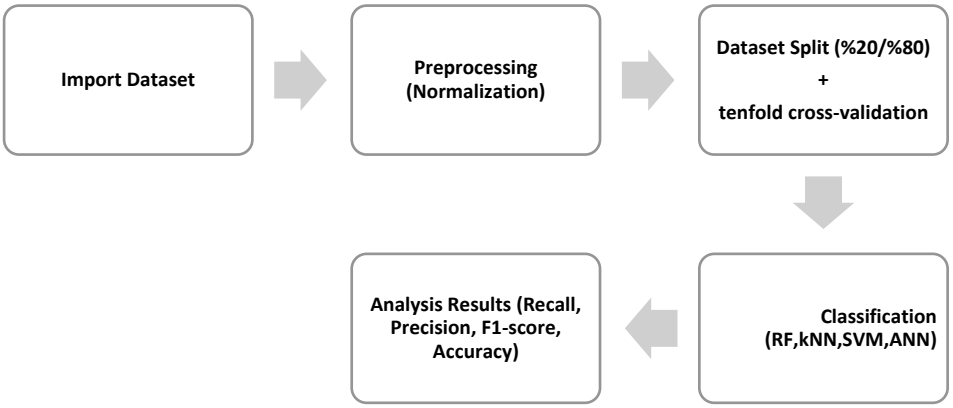

Fig. 1 – *Research methodology steps.*

The normalization was conducted using the StandardScaler method, adhering to the following formulas:

$$mean = \frac{\sum x}{n}, \tag{1}$$

$$std = \sqrt{\frac{\sum \left(x - mean \right)^2}{n-1}}, \tag{2}$$

$$x_{scaled} = \frac{x - mean}{std}. \tag{3}$$

Various models were employed to explore the dataset's underlying patterns for precise classification. These included the RF classifier, k-NN, SVM, and ANN. The ANN was constructed using the sequential model API from the Keras library with TensorFlow as the backend. Its architecture comprised an input layer, hidden layers employing the "ReLU" activation function, and an output layer utilizing "sigmoid" activation.

The training and evaluation of these models followed a rigorous process. Models were trained on the training subset, learning to map the relationships between features and the target variable. Subsequent evaluations on the testing subset assessed their predictive accuracy, robustness, and generalizability to unseen data. This comprehensive approach, encompassing diverse model architectures from the simplicity of k-NN to the complexity of ANN, facilitates a thorough understanding and comparison of different predictive methodologies. The findings from this study contribute significantly to developing robust and reliable predictive models for diabetes diagnosis.

3.3 Modeling and classification

Random Forest: The Random Forest (RF) classifier, an ensemble learning method introduced by Breiman [18], significantly improves prediction accuracy and robustness by integrating multiple Classification and Regression Trees (CARTs). This method employs a bagging technique where each tree is constructed using subsets of training samples drawn with replacement, as described by Belgiu and Drăguţ [19]. This process allows for the possibility of the same sample being selected multiple times or not in each subset.

This ensemble generates trees using random vectors independently sampled from the input vector. Each tree contributes unit votes for classifying an input vector, with the class receiving the most votes being selected as the final prediction [20]. The RF method randomly selects features at each decision split in the trees to reduce correlation, enhancing both predictive power and efficiency.

The utilization of RF mitigates the risk of overfitting, demonstrates resilience to outliers in the training data, and simplifies parameter setting, notably eliminating the need for tree pruning. As a result, RF retains the benefits inherent in Decision Trees and frequently surpasses them in performance. This enhancement is attributed to the method's robust voting scheme and capability to handle various subsets of variables effectively [18, 21].

K-Nearest Neighbors: The k-Nearest Neighbors (kNN) algorithm, a nonparametric method for classifying data, has been explored and described by several researchers, including Altman [22], Aha et al. [23], Ghosh et al. [24], and Jiang et al. [25]. It employs a straightforward two-step process: initially identifying the closest data points, termed 'neighbors,' and subsequently classifying a data point based on these neighbors. The proximity between data points is commonly calculated using the Euclidean distance formula: $D_{Euclid}(x,y)=\sqrt{\sum_i (x_i - y_i)^2}$, which measures the direct distance between points in the feature space.

In its operation, kNN classifies a data point by determining its 'k' nearest neighbors and then assigning it the most common class among them. This method exemplifies a lazy learning approach, where the learning process is deferred until the classification stage, unlike eager learning methods that involve constructing a predictive model during the training phase.

This unique mechanism endows kNN with significant value in classification tasks. It effectively leverages distance metrics to make accurate predictions, excelling in scenarios where the relational proximity of data points is a crucial determinant of their classification. The simplicity and effectiveness of kNN, which requires no assumptions about the underlying data distribution, make it a versatile tool for various predictive modeling applications.

Support Vector Machines: Support Vector Machines (SVM), initially conceptualized by Vapnik and colleagues in 1992 and further developed by Cortes and Vapnik [26] in 1995, are rooted in statistical learning theory. Using kernel methods, SVMs excel at creating optimal separation between two data classes, particularly in non-linear environments. Advanced by Vapnik in 1998, SVMs are recognized for their capability to discern complex patterns in diverse datasets, extending their applications from text categorization to image and handwriting digit recognition [27, 28]. The theoretical underpinnings of SVM, such as VC dimension and structural risk minimization, contribute to their scalability and flexibility across various domains. The versatility of SVMs is augmented by different kernel functions, facilitating the integration of prior knowledge into classification tasks, thus playing a pivotal role in their efficacy in both linear and non-linear classifications [28, 29].

At SVM's core is finding a hyperplane that best separates the data into classes. The equation w mathematically represents this concept $x + b = 0$ where w is the average vector to the hyperplane, x represents the data points, and b is the bias term. The objective is to maximize the margin between the data points of different classes, calculated as $2/\|w\|$.

The classification decision for a data point is based on the sign of the function $f(x) = \text{sgn}(wx + b)$. The choice of 'k' nearest neighbors determines the class assignment, with SVM classifying the data point based on the majority class among these neighbors. The versatility of SVM in handling both linear and non-linear data is significantly enhanced through the use of kernel functions, such as polynomial, RBF (radial basis function), and sigmoid, which transform the data into a higher-dimensional space where a linear separator is feasible.

SVM's strength lies in its ability to handle complex and high-dimensional data with high accuracy, making it a robust tool for classification tasks across various fields. The careful selection of kernel functions based on the dataset's characteristics is crucial in harnessing the full potential of SVM in both linear and non-linear classification scenarios.

Artificial Neural Network: Artificial Neural Networks (ANNs), advanced computational models that mimic the structure and function of natural nervous systems, are pivotal in diverse tasks such as classification, estimation, and detection. These networks, reflecting the human brain's complexity, consist of units termed neurons or nodes, which form a complex system when functioning together. Each node, acting as a computational unit, processes inputs with varying complexities, setting ANNs apart from traditional mathematical models due to their implicit ability to identify inter-parameter relationships. This characteristic significantly elevates their applicability in fields like classification and modeling [30, 31]. The architecture of ANNs, inspired by the human brain's interconnected

nerve cells, comprises input, hidden, and output layers, facilitating learning and adaptation through input and output data analysis, thus generating new input approximations – an invaluable process in engineering for solving complex problems [32, 33]. Central to ANNs are three primary components: architecture, learning algorithm, and activation function, where the learning algorithm optimizes the network's weights for peak performance [34], and the activation function maintains input-output relationship integrity. The learning process in ANNs involves a three-stage cycle of output calculation, error evaluation and correction, and weight adjustment, embodying a trial-and-error learning method [35]. Within each ANN, neurons analogous to human nerve cells comprise inputs (x_1, x_2, x_3,…, x_n), weights (w_1, w_2, w_3,…, w_n), a summing function, an activation function, and an output (y), collaboratively processing external information, thus underscoring the intricate and potent nature of ANNs in computational modeling [36].

Activation functions commonly used in artificial neural networks are as follows.

Rectified Linear Unit (ReLU) function: The ReLU function outputs the input directly if it is positive; otherwise, it outputs zero. This function is commonly used in hidden layers due to its computational efficiency.

$$f(x) = \max(0,\ x)\,. \tag{4}$$

Sigmoid Function: The Sigmoid function transforms real-valued numbers into a range between 0 and 1. It is often used in the output layer for binary classification problems.

$$f(x) = \frac{1}{1+\mathrm{e}^{-x}}\,. \tag{5}$$

Softmax function: The Softmax function is used for multi-class classification problems. It converts a vector of values into probabilities that sum up to 1, making it suitable for handling outputs where classification into multiple categories is required.

$$f(x_i) = \frac{\mathrm{e}^{x_i}}{\sum_j \mathrm{e}^{x_i}}\,. \tag{6}$$

Hyperbolic Tangent (Tanh) function: The Tanh function maps real numbers between –1 and 1. It is similar to the Sigmoid function but with a different value range.

$$f(x) = \tanh(x) = \frac{2}{1+\mathrm{e}^{-2x}} - 1\,. \tag{7}$$

Leaky ReLU (Leaky Rectified Linear Unit) function: Leaky ReLU is a variation of ReLU that allows a slight gradient (αx) when the input is negative. This helps mitigate the "dead neurons" problem with standard ReLU.

$$f(x) = \begin{cases} x, & \text{if } x > 0; \\ \alpha x, & \text{if } x \leq 0. \end{cases} \tag{8}$$

4 Results

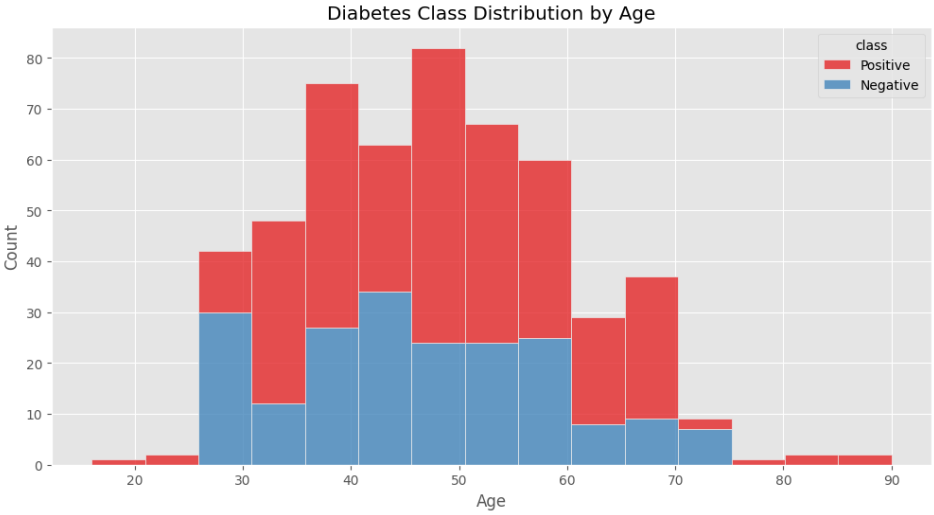

Fig. 2 – *Age distribution of patients.*

Fig. 2 illustrates the distribution of diabetes status by age, providing key insights into how the disease is concentrated across different age groups. The majority of positive diabetes cases are clustered in the 40 to 60 age range, with a noticeable peak around the ages of 50 and 55. This suggests that middle-aged adults are at a higher risk for developing diabetes, which corresponds with established risk factors such as age-related metabolic changes and lifestyle habits. Negative cases (non-diabetic) are more evenly distributed across the same age range but are fewer in comparison to positive cases. Both positive and negative cases are rare among younger individuals (under 30) and older adults (above 70). As shown in Fig. 2, these findings highlight the critical importance of early detection and intervention in middle-aged adults to prevent the onset and progression of diabetes. This age group should be a focal point for screening programs and targeted preventive healthcare measures.

Fig. 3 illustrates the distribution of diabetes status by gender, offering key insights into how the disease affects males and females. The graph shows that the majority of positive diabetes cases are found among females, with nearly equal numbers of male patients also diagnosed as positive. However, the number of negative cases (non-diabetic) is significantly higher among males compared to females. This suggests that while both genders are susceptible to diabetes, males

may have a higher rate of avoiding diagnosis or remaining undiagnosed compared to females. The noticeable difference in negative cases between the genders could indicate underlying socio-behavioral or biological factors influencing the diagnosis and prevalence of diabetes. These findings underscore the importance of considering gender as a significant factor in diabetes screening and prevention efforts.

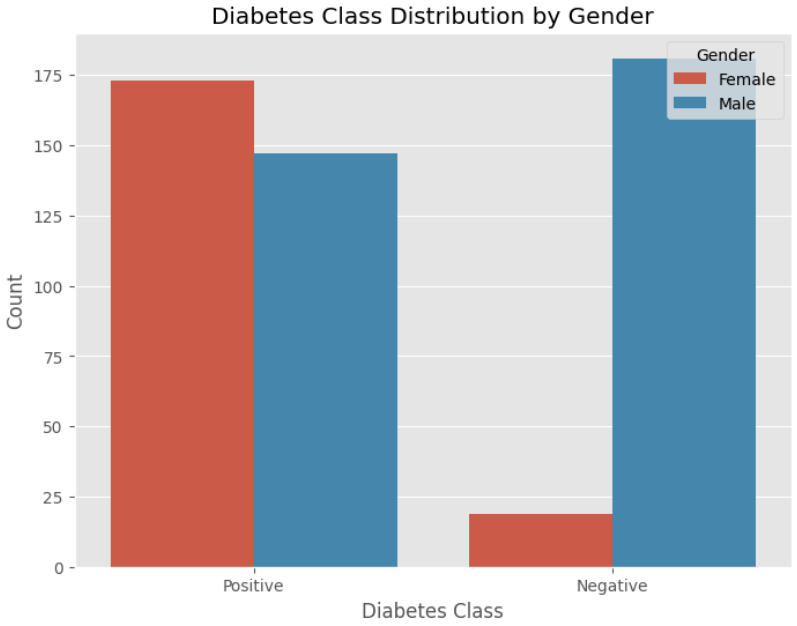

Fig. 3 – *Gender distribution of patients.*

In **Table 3**, the confusion matrices for four different classification models: ANN, RF, k-NN, and SVM—are meticulously represented. The confusion matrix is a critical tool in evaluating the efficacy of classification models, elucidating the instances of true positives, true negatives, false positives, and false negatives. The ANN model demonstrated commendable accuracy with 69 true positives and 32 true negatives but misclassified one as a false positive and two as false negatives. The RF model showcased exemplary precision, predicting 70 true positives and 33 true negatives, with only one false negative. The k-NN model, with 63 true positives and 31 true negatives, had a slightly higher misclassification rate, with two false positives and eight false negatives. The SVM model indicated balanced performance with 66 true positives and 28 true negatives but also had five false positives and false negatives. Analyzing these matrices, the Random Forest model emerges as the most precise, with minimal misclassifications, whereas the SVM model seems to have the highest misclassification rate.

Table 3
Analysis results of Confusion Matrix.

		ANN		RF		k-NN		SVM	
	Predicted								
	Class	0	1	0	1	0	1	0	1
Actual	0	32	1	**33**	**0**	31	2	28	5
	1	2	69	**1**	**70**	8	63	5	66

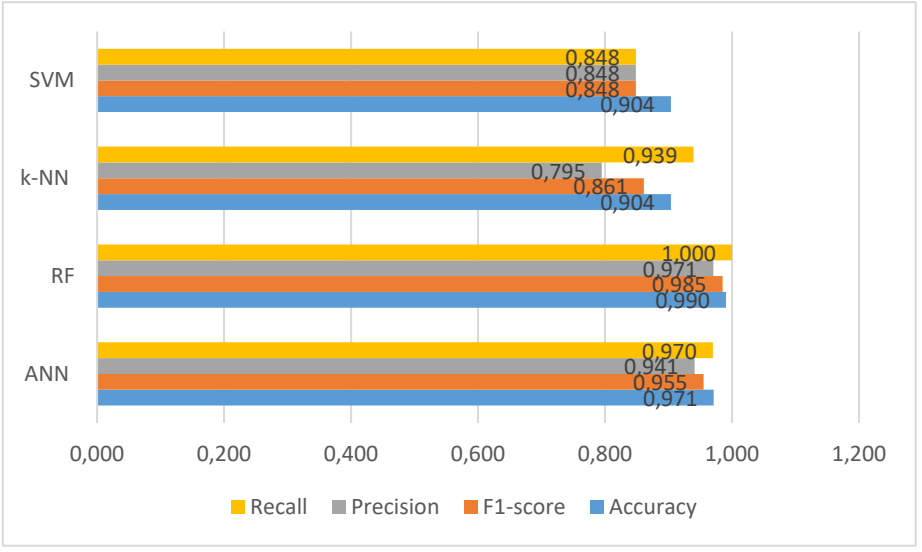

Fig. 4 – *Analysis results of the dataset using percentage split.*

The analytical results from the study present a comparative evaluation of four prominent machine learning models: ANN, RF, k-NN, and SVM. These results encapsulate the efficacy of each model in terms of Accuracy, F1-score, Precision, and Recall—a comprehensive set of metrics that collectively provide a holistic view of model performance.

The RF algorithm exhibits exceptional performance across all metrics, achieving the highest accuracy (0.990) and a perfect recall score (1.000). These figures indicate that the RF model not only correctly classified nearly all the instances but also managed to identify all the relevant cases as such. Furthermore, the RF model demonstrates a remarkable balance between precision (0.971) and sensitivity, as reflected by its F1-score (0.985), indicating high reliability and validity in its predictive capabilities.

In contrast, the ANN model also shows strong performance, with an accuracy of 0.971 and an F1-score of 0.955, complemented by precision (0.941) and recall (0.970) scores that are only marginally lower than those of the RF model. These results suggest that the ANN model is a robust classifier with a solid ability to generalize and effectively differentiate between classes.

The k-NN model, while demonstrating respectable accuracy (0.904), lags slightly behind in precision (0.795) and F1-score (0.861), with the recall rate at 0.939. These figures suggest that while k-NN is generally reliable, it may be prone to a higher rate of false positives, as indicated by its lower precision.

Lastly, the SVM model yields consistent scores across all metrics with accuracy, F1-score, precision, and recall at 0.904 and 0.848, respectively. While these figures reflect a balanced performance, they also highlight potential areas for improvement, particularly regarding the model's precision and recall balance.

In summary, the RF model outperforms the other models in this study, positioning it as a highly effective tool for predictive tasks in this context. The ANN follows closely, affirming its capacity as a powerful alternative. While effective, the k-NN and SVM models indicate a need for optimization to reach the performance levels of RF and ANN in diabetes classification tasks.

The analysis presented in Fig. 5 highlights the performance of four machine learning models: ANN, RF, k-NN, and SVM, evaluated through tenfold cross-validation. The results indicate notable variations in the models' effectiveness across the metrics of accuracy, F1-score, precision, and recall.

The RF model demonstrates superior performance, achieving the highest values across all metrics: an accuracy of 0.98, an F1-score of 0.98, a precision of 0.981, and a recall of 0.98. These results underscore the model's robustness in both identifying positive cases and minimizing false positives and false negatives, making it the most reliable classifier for this dataset.

The SVM model also performs strongly, with an accuracy of 0.923, an F1-score of 0.923, a precision of 0.926, and a recall of 0.923. While slightly lower than RF, SVM showcases high precision and recall, indicating its capability to balance between true positive rates and precision effectively.

The ANN model achieves an accuracy of 0.909, an F1-score of 0.91, a precision of 0.918, and a recall of 0.909. These results reflect ANN's strong predictive capacity, albeit slightly behind SVM and RF in overall performance.

In contrast, the k-NN model shows the lowest performance among the evaluated models, with an accuracy of 0.871, an F1-score of 0.872, a precision of 0.886, and a recall of 0.871. While k-NN demonstrates reasonable effectiveness, its performance is less competitive, suggesting it may be less suitable for datasets with this level of complexity.

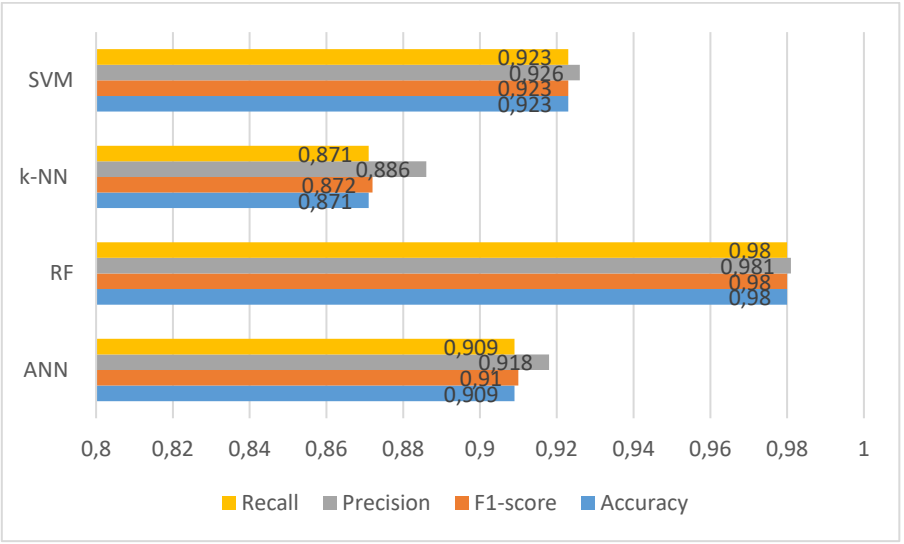

Fig. 5 – *Analysis results of the dataset using tenfold cross-validation.*

In summary, the results indicate that RF is the most effective model for this dataset, followed by SVM and ANN, with k-NN trailing. These findings highlight the importance of algorithm selection based on dataset characteristics and demonstrate the potential of RF and SVM for high-accuracy classifications in similar applications.

4.1 Feature importance

The attribute importance analysis conducted using the RF model has provided vital insights into the significance of each attribute in influencing the classification of diabetes. This analysis methodically ranked the attributes based on their contribution to the model's decision-making process, revealing key factors that impact classification outcomes.

Predominantly, Polyuria was identified as the most critical attribute, exerting the most significant influence on the classification. This attribute, indicative of excessive urination, is a primary symptom of diabetes. Hence, its prominence in the model aligns well with medical understanding. Following Polyuria, Polydipsia, which denotes intense thirst, was ranked as the second most influential attribute, reflecting its status as a cardinal symptom of diabetes.

Gender emerged as the third most significant attribute, suggesting the potential impact of biological differences in the manifestation or progression of diabetes. The fourth crucial attribute was Age, highlighting the role of age-related physiological changes in diabetes risk. Finally, sudden weight loss, another common symptom of diabetes, was identified as the fifth most crucial attribute. This underscores its relevance in the clinical presentation of the condition.

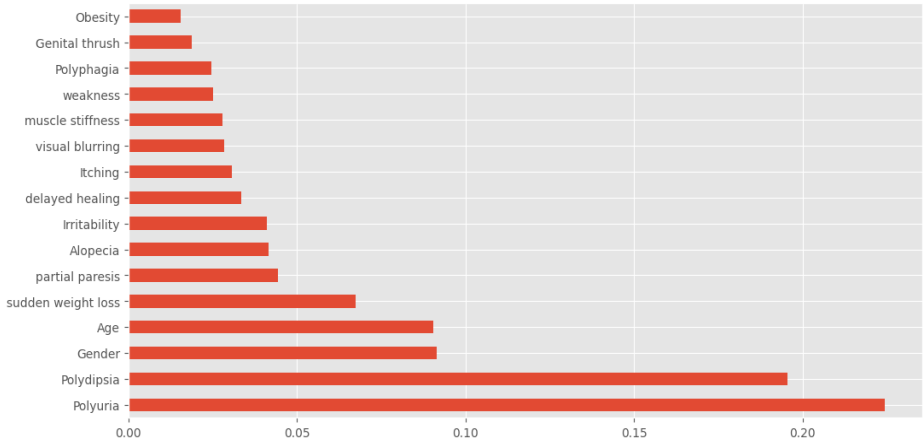

Fig. 5 – *Feature importance ranking with random forest algorithm.*

The hierarchical organization of these attributes, with Polyuria and Polydipsia at the forefront, followed by Gender, Age, and sudden weight loss, provides a nuanced understanding of the interplay between various factors in diabetes classification. This arrangement aligns with clinical symptoms and demographic influences and enhances our comprehension of the critical indicators in diabetes prediction.

These insights are invaluable for clinicians and researchers, aiding in early diagnosis and the formulation of tailored management strategies. Additionally, understanding the relative importance of these attributes can lead to more effective intervention development and improved risk assessment, ultimately contributing to enhanced patient outcomes and more informed healthcare approaches. By pinpointing the most influential factors in diabetes classification, this analysis paves the way for more targeted and efficient diagnostic procedures, facilitating better resource allocation and potentially leading to advancements in personalized medicine in diabetes care.

5 Discussion

This study explores the application of machine learning (ML) models for screening and assessing diabetes risk using patient-reported symptoms, addressing the global challenge of diabetes, particularly in low- and middle-income countries. As emphasized by the International Diabetes Federation (IDF), the increasing prevalence of diabetes demands innovative and scalable tools for early intervention [1]. By focusing on symptom-based data rather than clinical diagnostics, this research provides a non-invasive and accessible approach that aligns with global needs, especially in resource-constrained settings.

The inclusion of tenfold cross-validation ensures robust evaluation of the ML models, reducing biases associated with simple percentage splits and enhancing the reliability of the results. Among the tested models – Artificial Neural Networks, Random Forest, k-Nearest Neighbors, and Support Vector Machines – the RF model emerged as the most effective, achieving the highest metrics across all evaluation parameters: accuracy (0.98), precision (0.981), recall (0.98), and F1-score (0.98). These results align with prior findings, such as those reported by Tripathi and Kumar [11], Sarwar et al. [15], and Islam et al. [17], which also highlight RF's superior performance for similar datasets. RF's ensemble approach, which combines predictions from multiple decision trees, effectively minimizes overfitting and enhances predictive accuracy, particularly in high-dimensional datasets.

The SVM model, with an accuracy of 0.923 and a precision of 0.926, demonstrated balanced and reliable performance, making it a strong alternative to RF for diabetes risk screening. These findings are consistent with those of Joshi et al. [37], who identified SVM as a robust classifier for complex medical datasets. While SVM's performance was slightly lower than RF, its ability to maintain a balance between sensitivity and specificity underscores its utility in practical applications.

The ANN model performed well, achieving an accuracy of 0.909 and an F1-score of 0.91. ANNs are particularly adept at capturing non-linear and multifaceted patterns in data, as evidenced in prior studies such as Sapon et al. [38]. However, the slightly lower precision of ANN compared to RF highlights the trade-offs inherent in neural network-based approaches, which may require careful tuning to optimize performance.

In contrast, the k-NN model exhibited the lowest performance, with an accuracy of 0.871 and an F1-score of 0.872. Its proximity-based classification approach makes it prone to higher false positive rates, particularly in datasets with overlapping classes. These limitations suggest that while k-NN can serve as a baseline model, it is less suitable for real-world applications without further optimization or feature engineering.

The feature-important analysis conducted within the RF model offered valuable insights into the predictors of diabetes risk. Symptoms such as Polyuria and Polydipsia, already established as primary indicators of diabetes, emerged as the most significant features, corroborating clinical observations and findings from Padhi et al. [4, 5]. Additionally, demographic factors like age and gender were identified as critical predictors, reinforcing their role in diabetes epidemiology. The identification of sudden weight loss as a key feature further supports its inclusion in early screening frameworks, highlighting its clinical relevance in detecting early signs of diabetes. These findings are consistent with Islam et al. [17], who emphasized the importance of both common and less common symptoms for early detection through ML models.

In light of these findings, the results corroborate the potential of symptom-based screening tools as emphasized by Islam et al. [17], who demonstrated the utility of similar datasets for non-invasive risk prediction. Such tools can be instrumental in enabling early interventions, reducing the burden on healthcare systems, and improving patient outcomes. The insights from this study not only validate the applicability of RF and SVM as leading models for diabetes risk screening but also emphasize the importance of algorithm selection based on dataset characteristics. This work contributes to the growing body of research advocating for data-driven solutions in global health and underscores the necessity of scalable, accessible tools for addressing the diabetes epidemic effectively.

6 Conclusion

This study underscores the transformative potential of machine learning (ML) models in enhancing the early screening, diagnosis, and management of diabetes mellitus, a condition of growing global concern. By leveraging patient-reported symptoms and employing robust methodologies such as tenfold cross-validation, this research provides a comprehensive evaluation of ML models, with particular emphasis on their practical utility in healthcare systems, especially in resource-limited settings.

Among the evaluated models, Random Forest consistently demonstrated superior performance, achieving the highest accuracy and robustness across all evaluation metrics. Its exceptional predictive capability, coupled with its ability to perform feature importance analysis, positions RF as a cornerstone in diabetes diagnostics. These findings align with prior research by Tripathi and Kumar [11], Sarwar et al. [15], and Islam et al. [17], further validating RF's reliability for non-invasive diabetes screening and classification tasks.

The study also highlights the effectiveness of Artificial Neural Networks, which, although slightly trailing RF in performance, proved adept at capturing complex, non-linear relationships in the dataset. ANNs remain a valuable tool for medical diagnostics, particularly with further optimization. In contrast, while Support Vector Machines and k-Nearest Neighbors demonstrated respectable results, their performance suggests that additional refinement is required to maximize their clinical applicability.

The feature-important analysis within the RF model provided critical insights into key predictors of diabetes. Symptoms such as Polyuria and Polydipsia emerged as the most significant factors, corroborating clinical observations from Padhi et al. [4, 5]. Additionally, demographic factors like age and gender were highlighted as critical predictors, emphasizing their role in diabetes epidemiology. The identification of sudden weight loss further supports its relevance in early detection frameworks, underscoring the need to include these features in targeted screening programs.

Integrating ML models like RF into healthcare practices has the potential to revolutionize diabetes care by enabling scalable, non-invasive, and accurate screening tools. These models can help healthcare systems identify high-risk individuals earlier, facilitating timely interventions and reducing the global burden of diabetes. Future research should focus on refining these models, integrating them into real-world clinical workflows, and expanding their application to other chronic conditions. Furthermore, investigating the socio-behavioral and biological factors influencing model outcomes, particularly regarding gender disparities, could enhance the effectiveness of ML-driven healthcare solutions.

In conclusion, this study highlights the critical role of ML models in addressing the diabetes epidemic. By improving early detection and offering data-driven insights, these models pave the way for more personalized and effective healthcare strategies, ultimately improving the quality of life for millions affected by this chronic condition. Continued efforts to optimize and implement these technologies in diverse clinical settings will be pivotal in combating diabetes on a global scale.

7 References

[1] International Diabetes Federation, Facts & figures, Available at: https://idf.org/about-diabetes/diabetes-facts-figures [Accessed: 05-Dec-2023].

[2] World Health Organization, Diabetes, Available at: https://www.who.int/health-topics/diabetes [Accessed: 26-Dec-2023].

[3] G. Wilcox: Insulin and Insulin Resistance, The Clinical Biochemist. Reviews, Vol. 26, No. 2, May 2005, pp. 19 – 39.

[4] S. Padhi, A. K. Nayak, A. Behera: Type II Diabetes Mellitus: A Review on Recent Drug Based Therapeutics, Biomedicine & Pharmacotherapy, Vol. 131, November 2020, p. 110708.

[5] S. Padhi, M. Dash, A. Behera: Nanophytochemicals for the Treatment of Type II Diabetes Mellitus: A Review, Environmental Chemistry Letters, Vol. 19, No. 6, December 2021, pp. 4349 – 4373.

[6] D. L. Eizirik, L. Pasquali, M. Cnop: Pancreatic β-Cells in Type 1 and Type 2 Diabetes Mellitus: Different Pathways to Failure, Nature reviews. Endocrinology, Vol. 16, No. 7, July 2020, pp. 349 – 362.

[7] N. Katiyar, H. K. Thakur, A. Ghatak: Recent Advancements Using Machine Learning & Deep Learning Approaches for Diabetes Detection: A Systematic Review, e-Prime-Advances in Electrical Engineering, Electronics and Energy, Vol. 9, September 2024, p. 100661.

[8] M. Komi, J. Li, Y. Zhai, X. Zhang: Application of Data Mining Methods in Diabetes Prediction, Proceedings of the 2nd International Conference on Image, Vision and Computing (ICIVC), Chengdu, China, June 2017, pp. 1006 – 1010.

[9] E. Ramanujam, T. Chandrakumar, K. T. Thivyadharsine, D. Varsha: A Multilingual Decision Support System for Early Detection of Diabetes Using Machine Learning Approach: Case Study for Rural Indian People, Proceedings of the 5th International Conference on Research

in Computational Intelligence and Communication Networks (ICRCICN), Bangalore, India, November 2020, pp. 17 – 21.

[10] F. A. Khaleel, A. M. Al-Bakry: Diagnosis of Diabetes Using Machine Learning Algorithms, Materials Today: Proceedings, Vol. 80, No. 3, 2003, pp. 3200 – 3203.

[11] G. Tripathi, R. Kumar: Early Prediction of Diabetes Mellitus Using Machine Learning, Proceedings of the 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, June 2020, pp. 1009 – 1014.

[12] H. Kaur, V. Kumari: Predictive Modeling and Analytics for Diabetes Using a Machine Learning Approach, Applied Computing and Informatics, Vol. 18, No. 1/2, March 2022, pp. 90 – 100.

[13] J. Xue, F. Min, F. Ma: Research on Diabetes Prediction Method Based on Machine Learning, Journal of Physics: Conference Series, Vol. 1684, November 2020, p. 012062.

[14] D. Sisodia, D. S. Sisodia: Prediction of Diabetes Using Classification Algorithms, Procedia Computer Science, Vol. 132, 2018, pp. 1578 – 1585.

[15] M. A. Sarwar, N. Kamal, W. Hamid, M. Ali Shah: Prediction of Diabetes Using Machine Learning Algorithms in Healthcare, Proceedings of the 24th International Conference on Automation and Computing (ICAC), Newcastle Upon Tyne, UK, September 2018, pp. 1 – 6.

[16] V. Rawat, S. Joshi, S. Gupta, D. P. Singh, N. Singh: Machine Learning Algorithms for Early Diagnosis of Diabetes Mellitus: A Comparative Study, Materials Today: Proceedings, Vol. 56, No. 1, 2022, pp. 502-506.

[17] M. M. F. Islam, R. Ferdousi, S. Rahman, H. Y. Bushra: Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques, Proceedings of the 1st International Symposium on Computer Vision and Machine Intelligence in Medical Image Analysis (ISCMM), Sikkim Manipal Institute of Technology, India, February 2019, pp. 113 – 125.

[18] L. Breiman: Random Forests, Machine Learning, Vol. 45, No. 1, October 2001, pp. 5 – 32.

[19] M. Belgiu, L. Drăguț: Random Forest in Remote Sensing: A Review of Applications and Future Directions, ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 114, April 2016, pp. 24 – 31.

[20] M. Pal: Random Forest Classifier for Remote Sensing Classification, International Journal of Remote Sensing, Vol. 26, No. 1, 2005, pp. 217 – 222.

[21] J. Ali, R. Khan, N. Ahmad, I. Maqsood: Random Forests and Decision Trees, International Journal of Computer Science Issues, Vol. 9, No. 5, September 2012, pp. 272 – 278.

[22] N. S. Altman: An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, The American Statistician, Vol. 46, No. 3, August 1992, pp. 175 – 185.

[23] D. W. Aha, D. Kibler, M. K. Albert: Instance-Based Learning Algorithms, Machine Learning, Vol. 6, No. 1, January 1991, pp. 37 – 66.

[24] R. Ghosh, S. Phadikar, N. Deb, N. Sinha, P. Das, E. Ghaderpour: Automatic Eyeblink and Muscular Artifact Detection and Removal from EEG Signals Using k-Nearest Neighbor Classifier and Long Short-Term Memory Networks, IEEE Sensors Journal, Vol. 23, No. 5, March 2023, pp. 5422 – 5436.

[25] L. Jiang, Z. Cai, D. Wang, S. Jiang: Survey of Improving k-Nearest-Neighbor for Classification, Proceedings of the 4th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Haikou, China, August 2007, pp. 1 – 5.

[26] C. Cortes, V. Vapnik: Support-Vector Networks, Machine Learning, Vol. 20, No. 3, September 1995, pp. 273 – 297.

[27] V. N. Vapnik: Statistical Learning Theory, Wiley, New York, 1998.

[28] N. Cristianini, J. Shawe-Taylor: An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, 2000.

[29] A. M. Shahiri, W. Husain, N. A. Rashid: A Review on Predicting Student's Performance Using Data Mining Techniques, Procedia Computer Science, Vol. 72, 2015, pp. 414 – 422.

[30] A. Jahanbakhshi, R. Salehi: Processing Watermelon Waste Using Saccharomyces Cerevisiae Yeast and the Fermentation Method for Bioethanol Production, Journal of Food Process Engineering, Vol. 42, No. 7, November 2019, p. e13283.

[31] K. Utai, M. Nagle, S. Hämmerle, W. Spreer, B. Mahayothee, J. Müller: Mass Estimation of Mango Fruits (Mangifera indica L., cv. 'Nam Dokmai') by Linking Image Processing and Artificial Neural Network, Engineering in Agriculture, Environment and Food, Vol. 12, No. 1, January 2019, pp. 103 – 110.

[32] K. Öztürk, M. E. Şahin: A General View of Artificial Neural Networks and Artificial Intelligence, Takvim-i Vekayi, Vol. 6, No. 2, December 2018, pp. 25 – 36.

[33] B. Yegnanarayana: Artificial Neural Networks, PHI Learning Pvt. Ltd, Delhi, 2009.

[34] D. Graupe: Principles of Artificial Neural Networks, 3rd Edition, World Scientific Publishing Co. Pte. Ltd., New Jersey, London, Singapore, 2013.

[35] D. J. Livingstone: Artificial Neural Networks: Methods and Applications, Humana Press, Totowa, 2008.

[36] S. Haykin: Neural Networks: A Comprehensive Foundation, 2nd Edition, Pearson Education (Singapore) Pte. Ltd., Delhi, 1999.

[37] T. N. Joshi, P. M. Chawan: Diabetes Prediction Using Machine Learning Techniques, International Journal of Engineering Research and Application, Vol. 8, No. 1, January 2018, pp. 9 – 13.

[38] M. A. Sapon, K. Ismail, S. Zainudin: Prediction of Diabetes by Using Artificial Neural Network, Proceedings of the International Conference on Circuits, System and Simulation (ICCSS), Bangkok, Thailand, May 2011, pp. 299 – 303.