



Comprehensive Feature Selection Methods for Predicting Diabetes and Stroke Risk

Karthik Balaji¹ and Alexander Iliev Iliev^{1,2}(✉)

¹ SRH University of Applied Sciences, Berlin, Germany
ailiev@berkeley.edu

² Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

Abstract. The study titled “Comprehensive Feature Selection Methods for Predicting Diabetes and Stroke Risk Among Married Individuals” explores the relationship between various risk factors and the prevalence of diabetes and stroke among married individuals. The primary aim is to identify key risk factors associated with diabetes and stroke among married, divorced, or widowed individuals. To achieve this, a robust feature selection process is employed to identify the most influential factors among an extensive set of variables. After a thorough literature review, the project proceeds with data preparation. This involves the selection of 21 relevant features based on domain knowledge and data analysis. Additionally, the dataset undergoes a comprehensive cleaning process to ensure data quality and accuracy. To select the most relevant features for diabetes and stroke the cleaned dataset was made to undergo different feature selection methods including some new and innovative methods along with some well-known methods. These features were trained using different machine learning algorithms and classifiers and the most relevant features affecting diabetes and stroke were evaluated based on the performance metrics. Prediction models for both diabetes and stroke were then built using the most relevant features. The innovative methodology, integrating various feature selection techniques, advances the field by enhancing the accuracy and robustness of predictive models. The research findings open doors for further investigations and provide a significant contribution to the healthcare domain.

Keywords: Risk factors · Diabetes · Stroke · Machine learning · Feature selection · BRFSS data

1 Introduction

Diabetes [1] and stroke represent significant health challenges with profound implications for individuals and societies worldwide. This study embarks on a comprehensive inquiry to explore the risk factors influencing diabetes and stroke within married populations. Leveraging advanced data analysis techniques and machine learning methodologies, this study undertakes a rigorous examination of the multifaceted aspects of diabetes and stroke risk. The urgency of this research is underscored by the critical need to comprehend the intricate interplay of factors contributing to diabetes and stroke occurrences among married individuals. This unique subset of the population demands

special attention due to potential distinct risk profiles. By uncovering these risk factors and potential preventive strategies, this study aims to contribute to evidence-based healthcare interventions tailored to the needs of this group. Focusing on the Behavioural Risk Factor Surveillance System (BRFSS) 2021 dataset, a valuable repository of health-related information, this research investigates the health dynamics of married populations. By employing feature selection techniques and predictive modelling, the study strives to discern the most influential risk factors for diabetes and stroke among married individuals.

1.1 Research Gap

While extensive research has explored risk factors for diabetes and stroke, there is a significant dearth of studies that specifically investigate these factors within the context of married populations. While spousal diabetes has been examined as a potential risk factor, comprehensive studies of risk factors for both diabetes and stroke within married populations are notably absent from the literature. Feature selection methodologies are pivotal in machine learning for identifying the most relevant attributes for model construction. However, a comprehensive evaluation of various features selection techniques and their effectiveness for diabetes and stroke risk prediction in married populations is lacking. This research aims to address these gaps by examining the risk factors for diabetes and stroke within the married population and proposing innovative feature selection methods.

1.2 Objective and Research Questions

The primary objective of this research is to develop accurate predictive models for assessing diabetes and stroke risks among married individuals by employing advanced data analysis techniques and feature selection methods. The overarching goals include:

- **Conducting Comprehensive Feature Selection:** By exploring an array of feature selection techniques, this research aims to develop a robust methodology for identifying the most pertinent risk factors [3].
- **Justifying Methodological Choices:** Each selected feature selection method serves a unique purpose in the context of diabetes and stroke risk assessment. This research will offer clear rationales for their selection.
- **Evaluating Performance:** Rigorous evaluation of feature selection methods and machine learning algorithms will be conducted to ensure accuracy and computational efficiency.
- **Assessing Applicability:** This research ensures that the chosen methods are not only accurate but also practically applicable to healthcare interventions.

To guide this investigation, several research questions have been formulated:

- a) *What are the prominent risk factors associated with diabetes and stroke in married populations?*
- b) *How effective are various feature selection methods, such as PCA, Boruta, LASSO, Chi-Square, Correlation, and Mutual Information, in identifying relevant risk factors?*

- c) *How does Integrated Feature Selection (IFS), Hybrid Feature Selection (Hybrid FS), Deep Feature Selection (Deep FS), Genetic Feature Selection (Genetic FS) contribute to risk factor identification for married populations?*
- d) *What are the most accurate predictive models for diabetes and stroke risk assessment among married individuals?*

By addressing these research questions, this study aims to advance our understanding of diabetes and stroke risk factors within married populations and provide valuable insights for evidence-based healthcare interventions tailored to this specific group. The intersection of machine learning techniques and healthcare has yielded significant advancements in the development of predictive models for various health conditions. This literature review delves into research contributions that utilize machine learning to construct risk prediction models, with a specific focus on predicting diabetes and stroke risks among married male and female individuals. By exploring feature selection methodologies across these studies, the aim is to enhance the precision and interpretive capacity of the models.

The studies covered in this review collectively underscore the vital connection between machine learning methodologies and healthcare prediction. Though not all are directly tailored for diabetes and stroke risk prediction, they offer insights applicable to our context. For instance, emphasizes the significance of behavioural factors in cardiovascular health prediction, setting the foundation for relevant feature selection techniques. Some highlight the integral role of physical activity and lifestyle in diabetes/stroke risks, guiding our feature selection process. One research provides a panoramic view of machine learning and deep learning models for type 2 diabetes prediction, potentially influencing our feature selection. The advanced feature selection methods in some research align with our comprehensive approach. Furthermore, a very few research elucidate diverse feature selection methods, crucial for refining our predictive model. Some contribute to understanding tailored approaches for health risk prediction, while few papers directly address diabetes prediction, providing methodological guidance. The amalgamation of insights from these studies equips us with a robust foundation for our proposed predictive model. This model is envisioned to be highly personalized, accurately assessing the risks of diabetes and stroke among married individuals through a meticulous feature selection process.

2 Methodologies

This methodology chapter serves as a strategic roadmap to address a critical research gap in health prediction. The aim is to create a personalized predictive model for married individuals, specifically targeting diabetes and stroke risks. The absence of such models for this demographic underscores the importance of this methodology. Conventional risk assessment methods often overlook the complex interplay of factors affecting married individuals, including shared lifestyle behaviors and genetics. The chosen methods bridge existing research gaps. Feature selection methods, informed by previous studies, ensure the inclusion of the most relevant variables. This methodology details the intricacies of data collection, preprocessing, and model integration. It emphasizes transparency and precision, making each step clear for replication and validation by other researchers.

By establishing this robust foundation, the research gains credibility and impact. This chapter invites readers to embark on a journey where scientific rigor and innovative techniques drive accurate and personalized health risk predictions.

2.1 Data Cleaning and Pre-processing

For this research, information was sourced from the *Behavioral Risk Factor Surveillance System (BRFSS)* 2021 dataset, provided by the *Centers for Disease Control and Prevention (CDC)* in the United States. This dataset serves as a comprehensive repository of health-related information accrued thru surveys, making it especially suitable for analyzing the prediction of diabetes and stroke danger amongst married male and lady individuals. The raw BRFSS 2021 dataset, comprising 193,993 records and 21 features, underwent meticulous data cleaning and preprocessing to ensure quality and reliability. The following steps were undertaken as shown in Table 1.

- 1) **Filtering by Marital Status:** Only records pertaining to married individuals were included, focusing the study on the specified population subset. Marital values greater than 4 were excluded.
- 2) **Handling Missing Values:** Records with missing values were removed using the 'dropna()' function, contributing to a more refined dataset.
- 3) **Recoding and Transformation:** Several features were recoded and transformed for consistency and meaningful interpretation. For instance, values representing health conditions were replaced or removed based on ethical considerations.
- 4) **Variable Name Adjustments:** Variable names were updated to enhance their intuitiveness and clarity, aligning with the underlying health indicators.
- 5) **Data Structure Preservation:** Categorical, ordinal, and continuous features were carefully handled to retain their inherent information structure.

The culmination of these steps led to the creation of the 'Cleaned_brfs.csv' dataset, featuring 21 features and 193,993 records. The data cleaning process as shown in Table 1 ensured data quality, integrity, and readiness for accurate predictive modelling of diabetes and stroke risks among married male and female individuals.

2.2 Feature Selection Methods Overview

Continuing from the previous section, the choice of appropriate feature selection methods is pivotal for enhancing the effectiveness of predictive models. This section outlines key feature selection methods that can be applied to the cleaned dataset to identify the most relevant and influential features for diabetes and stroke risk prediction among married male and female individuals [7].

Wrapper Methods, Embedded Methods, Filter Methods, Dimensionality Reduction Techniques. Each of these methods possesses its own strengths and limitations, making their selection context dependent. By applying and comparing multiple methods, we aim to identify the most influential features contributing significantly to diabetes and stroke risk prediction. Through a systematic feature selection process, we aim to enhance the model's accuracy, interpretability, and generalizability within the specified population. IFS combines multiple methods for robust selection, HFS balances

Table 1. Data pre-processing and cleaning

ORIGINAL FEATURE	<i>CORRESPONDING QUESTION IN BRFSS SURVEY</i>	CHANGES MADE
DIABETE4	Have you ever been told you have diabetes?	Removed values: 7 (Don't know), 9 (Refused). Replaced values: 1, 3, 4 with 0 (No)
_MICH	Ever been told you have coronary heart disease?	Replaced value: 2 with 0 (No). Removed values: 7 (Don't know), 9 (Refused)
TOLDHI3	Ever been told cholesterol is high?	Replaced values: 2, 3 with 0 (No). Removed value: 9 (Refused)
_BMI5	Body Mass Index	Divided values by 100 to convert to real BMI values
SMOKE100	Have you smoked at least 100 cigarettes in life?	Removed values: 7 (Don't know), 9 (Refused). Replaced value: 2 with 0 (No)
_CHOLCH3	Cholesterol check within past five years	Replace values 2 and 3 with 0; Remove 9
CVDSTRK3	Ever had a stroke?	Replaced value: 2 with 0 (No). Removed values: 7 (Don't know), 9 (Refused)
TOTINDA	Physical Activity in the past 30 days	Replaced value: 2 with 0 (No). Removed value: 9 (Don't know)
_FRTLT1A	Do you consume 1 or more fruits per day?	Replaced value: 2 with 0 (No). Removed value: 9 (Refused)
_VEGLT1A	Do you consume 1 or more vegetables per day?	Replaced value: 2 with 0 (No). Removed value: 9 (Refused)
_RFDRHV7	Heavy Alcohol Consumption	Replaced value: 1 with 0 (No), 2 with 1 (Yes). Removed value: 9 (Refused)
_HLTHPLN	Do you have any kind of health care coverage?	Replaced value: 2 with 0 (No). Removed values: 7 (Don't know), 9 (Refused)
MEDCOST1	Medical cost prevents seeking necessary care	Replaced value: 2 with 0 (No). Removed values: 7 (Don't know), 9 (Refused)
GENHLTH	General Health Condition	Removed values: 7 (Don't know), 9 (Refused)

(continued)

Table 1. (continued)

ORIGINAL FEATURE	CORRESPONDING QUESTION IN BRFSS SURVEY	CHANGES MADE
MENTHLTH	Mental Health Condition in the past 30 days	Replaced value: 88 with 0 (No bad mental health days). Removed values: 77 (Don't know), 99 (Refused)
RFHYPE6	Told by a doctor or nurse if you have high BP?	Replaced value: 2 with 0 (No). Removed values: 7 (Don't know), 9 (Refused)
PHYSHLTH	Physical Health Condition in the past 30 days	Replaced value: 88 with 0 (No bad physical health days). Removed values: 77 (Don't know), 99 (Refused)
DIFFWALK	Serious difficulty walking or climbing stairs	Replaced value: 2 with 0 (No). Removed values: 7 (Don't know), 9 (Refused)
SEXVAR	Respondent's Gender	Replaced value: 2 with 0 (Female)
AGEG5YR	Age Group	Removed value: 14 (Don't know or missing)
EDUCA	Highest Grade or Year of School Completed	Removed value: 9 (Refused)
INCOME3	Annual Household Income from all sources	Removed values: 77 (Don't know), 99 (Refused)

filter and wrapper techniques, GFS evolves feature subsets efficiently, and DLFS leverages deep learning for automated, interpretable selection. Each method brings unique strengths to your research, offering a comprehensive toolkit for predicting diabetes and stroke risk among married individuals [4, 11].

Integrated Feature Selection (IFS) is an innovative approach designed to enhance feature selection for predictive modelling. The IFS process involves individual feature selection using various methods, followed by a voting system that aggregates selections from different methods. IFS addresses limitations of using individual methods by providing enhanced robustness, handling diverse data distributions, and offering a consensus-based approach for more accurate and stable feature selection.

Hybrid Feature Selection (HFS) combines filter (PCA) and wrapper (Boruta, LASSO) methods to select relevant features. It is a balanced method that capitalizes at the strengths of both paradigms. HFS involves training a deep neural network model and using gradient-based feature importance to select influential features. While computationally demanding, DLFS automatically learns informative features, providing a data-driven and comprehensive approach to selection.

Genetic Feature Selection (GFS) is inspired by natural selection, using genetic operators like crossover and mutation to evolve feature subsets. It's adaptable, captures interactions among features, and automates feature selection, reducing the need for manual engineering. GFS's ability to optimize feature selection iteratively enhances model performance with fewer features, leading to faster training and real-time prediction. Its focus on reducing redundancy and improving interpretability makes GFS invaluable for healthcare applications.

Deep Learning Feature Selection (DLFS) leverages deep neural networks to automatically learn relevant features. While computationally intensive, DLFS offers an automated and efficient approach, reducing the need for manual selection. DLFS's power lies in its ability to uncover intricate relationships, potentially mitigating overfitting, and creating interpretable feature subsets.

Each of these methods possesses its own strengths and limitations, making their selection context dependent. By applying and comparing multiple methods, we can identify the most influential features that contribute significantly to diabetes and stroke risk prediction. IFS combines multiple methods for robust selection, HFS balances filter and wrapper techniques, GFS evolves feature subsets efficiently, and DLFS leverages deep learning for automated, interpretable selection.

2.3 Machine Learning Algorithms Overview

In this study, a comprehensive assessment of ten machine learning algorithms has been conducted to predict the risk of diabetes and stroke among married individuals. These algorithms were selected to offer a diverse representation of both conventional and contemporary strategies, allowing a robust evaluation of predictive performance across various methodologies. The algorithms under scrutiny are as follows: Naive Bayes, Extra Trees, XGBoost, Gradient Boost, LightGBM, AdaBoost, Random Forest, Logistic Regression, MLP Classifier, and CatBoost [5].

Extra Trees is an ensemble method that builds multiple decision trees and aggregates their predictions. **XGBoost** is a popular gradient boosting algorithm known for its strong predictive power. It constructs a series of decision trees and iteratively corrects errors made by preceding trees, thereby boosting performance. **Gradient Boosting** constructs an ensemble of decision trees sequentially, with each tree aiming to correct the mistakes of the previous ones. **LightGBM** is a gradient boosting framework that emphasizes efficiency and speed. **Random Forest** constructs multiple decision trees using bootstrapped datasets and features. The algorithm averages predictions from all trees, reducing overfitting and improving generalization. **CatBoost** is a gradient boosting algorithm that handles categorical features effectively. It uses ordered boosting and provides robust performance with minimal hyperparameter tuning [8].

Each of these algorithms was evaluated on its ability to predict diabetes and stroke risk among married individuals. Performance metrics such as accuracy, precision, recall, F1-score, and area under the curve (AUC) were used to assess their predictive capabilities. The inclusion of a diverse set of algorithms ensured a comprehensive understanding of their respective strengths and weaknesses in addressing the research problem. By providing clear insights into our research design, data collection, preprocessing, feature selection, and machine learning algorithms, we lay the groundwork for advancing the

field of health risk prediction and contributing to the overarching goal of personalized healthcare interventions [12, 13].

3 Findings

Through a meticulous examination of feature selection methods for predicting diabetes and stroke risk among married individuals, several significant findings have emerged, contributing to a deeper understanding of the predictive attributes and their implications for healthcare interventions [9].

3.1 Feature Selection Methods Analysis

The evaluation of feature selection methods revealed varying computational efficiencies and relevance. Genetic Feature Selection (FS) was notably time-consuming, while methods like Correlation and Chi-square were swift. This underscores the need to balance feature accuracy with computational efficiency. For resource-constrained situations, Correlation and Chi-square prove advantageous. Across diabetes and stroke risk prediction, certain attributes consistently emerged as indicators, such as ‘HighBP’, ‘HighChol’, ‘Smoker’, and ‘HeartDiseaseorAttack’, reinforcing their medical relevance. Lifestyle attributes like ‘PhysActivity’, ‘Fruits’, and ‘Veggies’ highlight the impact of personal habits on risk prediction [14, 15]. The Integrated Feature Selection method excels in comprehensive attribute consideration. Genetic FS, despite being time-intensive, unveils a rich feature set for intricate variable interactions. These findings have significant implications for healthcare. By incorporating these features into predictive models, personalized interventions can be developed. This tailored approach, combining medical history, lifestyle factors, and pertinent indicators, holds potential for reducing the incidence of diabetes and strokes through targeted preventive measures.

Diabetes Prediction. The analysis of machine learning algorithms for diabetes prediction yielded valuable insights into their performance and suitability for this critical healthcare task. The comparison of ten algorithms using accuracy, ROC-AUC, and runtime metrics revealed distinct patterns and trade-offs [10, 18].

The standout performer across accuracy and computational efficiency was **LightGBM**. Its optimized gradient boosting implementation enabled it to consistently achieve superior accuracy compared to other models while maintaining acceptable response times. Genetic feature selection emerged as the top method for maintaining model consistency and relevance. However, Naive Bayes and Extra Trees exhibited suboptimal performance due to their limitations in handling complex dependencies and randomness in feature selection, respectively. The selection of LightGBM over Gradient Boosting was grounded in LightGBM’s histogram-based approach, which facilitated faster training without compromising accuracy. The comparative evaluation provided a comprehensive understanding of the strengths and weaknesses of each algorithm, empowering decision-makers to make informed choices for diabetes prediction tasks. Our findings recommend **LightGBM** with *genetic feature selection* (see Figs. 1, 2, 3, 4 and 5) as the optimal combination for this predictive model, considering both accuracy and computational efficiency as we can see from Table 2 and Table 3 respectively.

Table 2. Runtime evaluation (Diabetes)

Feature Selection Methods	Time (in seconds)
PCA	2.38
Boruta	146.07
LASSO	1.32
Correlation	0.7
Chi-square	0.067
Mutual Information	29.3
Integrated FS	128.37
Hybrid FS	112.06
Deep FS	272.88
Genetic FS	4001.93

Table 3. Performance evaluation (Diabetes)

Algorithm	Accuracy	ROC-AUC	Runtime
Naive Bayes	0.7647 (± 0.0046)	0.7720 (± 0.0059)	1.0001 s
Extra Trees	0.8339 (± 0.0023)	0.7318 (± 0.0041)	134.8407 s
XGBoost	0.8538 (± 0.0011)	0.8030 (± 0.0046)	33.9785 s
Gradient Boost	0.8555 (± 0.0013)	0.8083 (± 0.0041)	142.4828 s
LightGBM	0.8557 (± 0.0012)	0.8080 (± 0.0041)	5.3626 s
ADABOOST	0.8547 (± 0.0010)	0.8049 (± 0.0045)	49.8157 s
RANDOM FOREST	0.8404 (± 0.0018)	0.7587 (± 0.0030)	136.8348 s
LOGISTIC REGRESSION	0.8542 (± 0.0013)	0.8021 (± 0.0049)	1.7485 s
MLP CLASSIFIER	0.8510 (± 0.0026)	0.7954 (± 0.0057)	888.2004 s
Catboost	0.8539 (± 0.0009)	0.8057 (± 0.0044)	167.5063 s

The selection of relevant features plays a pivotal role in enhancing the model’s accuracy and efficiency. The chosen features include a range of factors such as *‘HighBP,’ ‘HighChol,’ ‘CholCheck,’ ‘BMI,’ ‘Stroke,’ ‘HeartDiseaseorAttack,’ ‘HvyAlcoholConsump,’ ‘DiffWalk,’ ‘Age,’ ‘Income,’ ‘PhysHlth,’ ‘MentHlth,’ ‘GenHlth,’ ‘Education,’ and ‘Fruits’* [16]. Comparing the Figs. 1, 2, 3, 4 and 5 we can infer that the *LightGBM* classifier takes the center stage in the model’s architecture and its performance is redefined through hyperparameter tuning. The resultant accuracy achieved by this model for diabetes prediction stands at approximately **85.59%**, accompanied by a runtime of **65.96** s. This trained model becomes a valuable tool for diabetes prediction and further clinical applications. The integration of the trained diabetes prediction model with

user interaction and personalized recommendations paves the way for a comprehensive approach towards proactive health management. The model, rooted in sophisticated techniques like Genetic Feature Selection and LightGBM, leverages users’ inputs to calculate a diabetes risk score. This score, which signifies the likelihood of diabetes development, offers a clear perspective on one’s health status. By interpreting the risk score, personalized recommendations are generated, aligning with evidence-based healthcare guidelines. These recommendations cater to individuals with varying risk scores, guiding them towards specific actions that align with their risk profile. The recommendations are meticulously justified, grounded in medical research and established correlations between lifestyle choices and diabetes risk [17, 20].

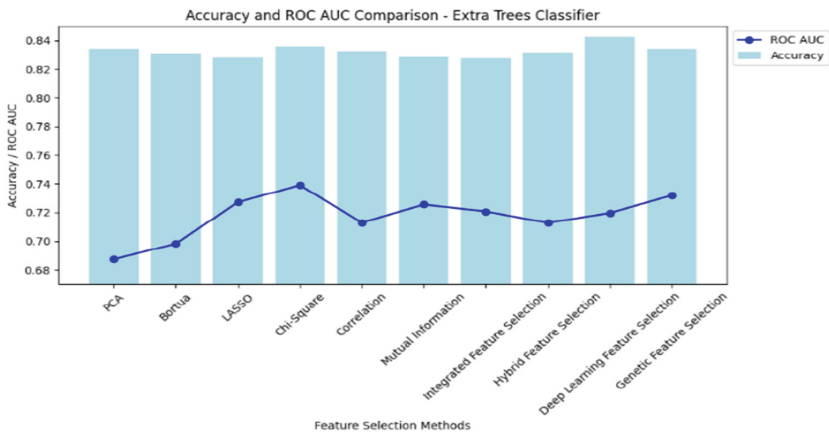


Fig. 1. Extra trees classifier

Stroke Prediction. The examination of machine learning algorithms for stroke prediction shed light on their effectiveness in tackling this significant medical challenge. The assessment of ten algorithms using accuracy, ROC-AUC, and runtime metrics presented valuable insights into their performance and trade-offs.

Our findings recommend LightGBM with integrated feature selection as the optimal choice for this predictive model, offering a balance between accuracy and computational efficiency.

In conclusion, our research underscores the significance of advanced feature selection techniques in enhancing the predictive capabilities of the developed models, thereby providing more reliable and actionable insights for both diabetes and stroke prediction. The selected algorithms’ performances highlight the intricate interplay between accuracy and computational efficiency, aiding healthcare professionals and policymakers in making informed decisions for early disease diagnosis and management. In this phase, we focus on training the stroke prediction model using Integrated Feature Selection (IFS). The stroke prediction model achieves an impressive accuracy of approximately 95.79% as we can see from Table 5 and it operates efficiently, with a runtime of 239.17 s as we can see from Table 4. The process of training the stroke prediction model and integrating it with user interaction and personalized recommendations mirrors that of

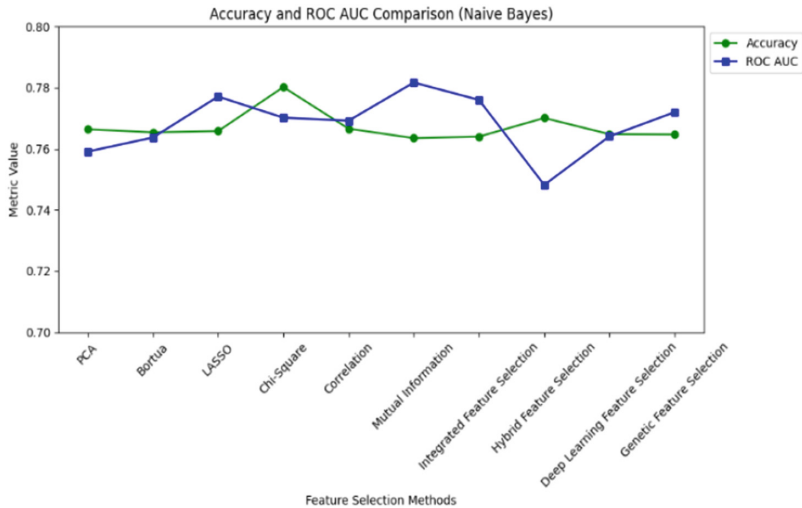


Fig. 2. Naive Bayes'

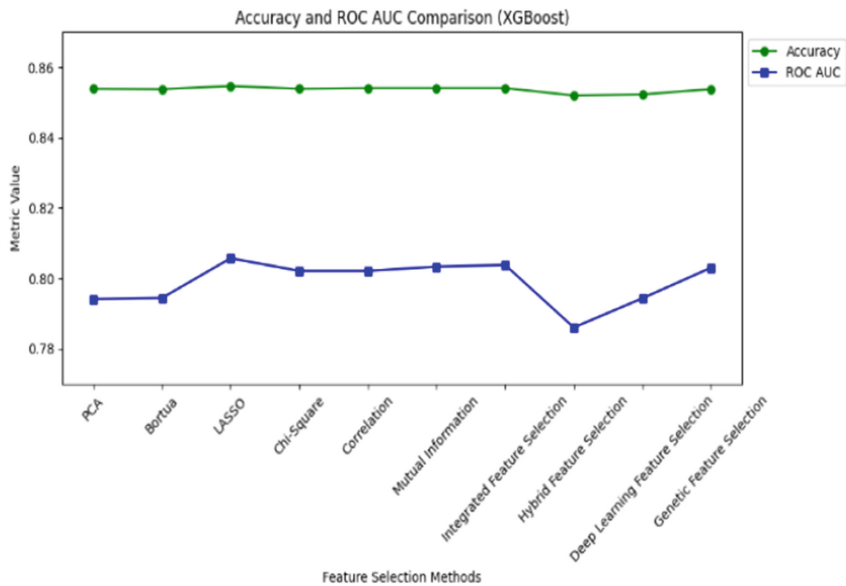


Fig. 3. XG boost classifier

the diabetes prediction model. The advanced stroke risk prediction model, empowered by Integrated Feature Selection and LightGBM, calculates an individual’s stroke risk score based on user inputs.

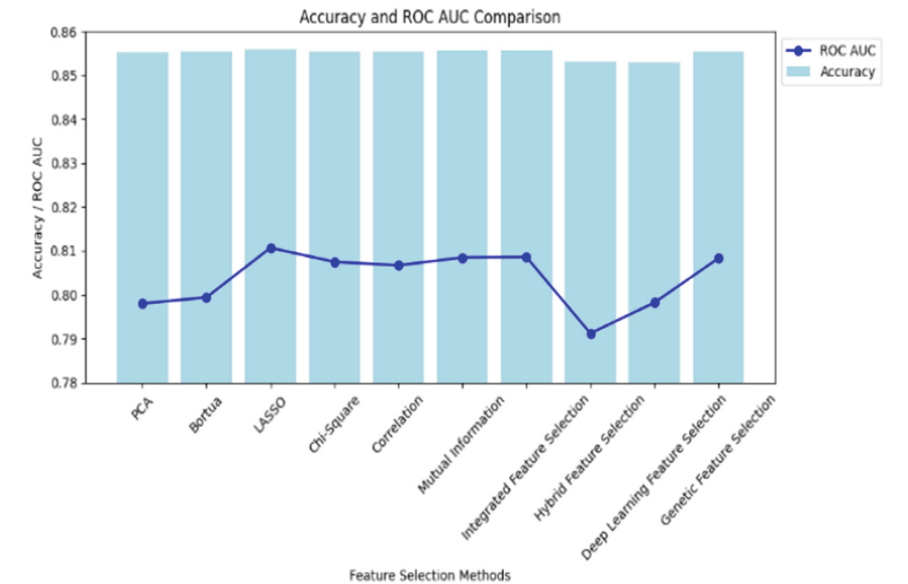


Fig. 4. Gradient boosting

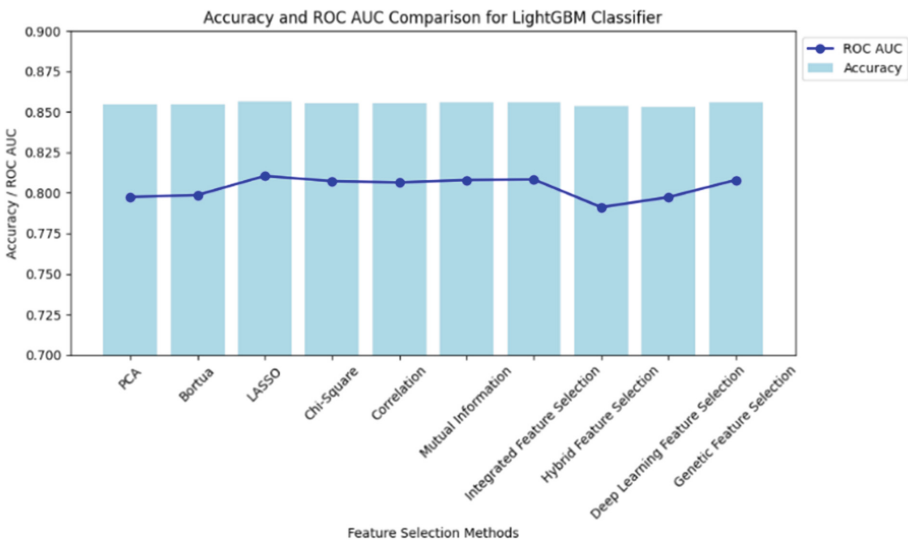


Fig. 5. LightGBM classifier

The findings of this research paper underscore the significance of advanced machine learning techniques in healthcare predictive modelling. Through rigorous training, hyperparameter tuning, and feature selection methods like Genetic Feature Selection and Integrated Feature Selection, the developed models demonstrated remarkable predictive

Table 4. Runtime evaluation (Stroke)

Feature Selection Methods	Runtime (in seconds)
PCA	1.32
Boruta	89.47
LASSO	1.27
Correlation	0.59
Chi-square	0.06
Mutual Information	27
Integrated FS	92.58
Hybrid FS	99.49
Deep FS	213.36
Genetic FS	2556.93

Table 5. Performance evaluation (Stroke)

Algorithm	Accuracy	ROC-AUC	Runtime (sec)
Naive Bayes	0.8587 (± 0.0026)	0.7918 (± 0.0084)	0.83
Extra Trees	0.9513 (± 0.0007)	0.6985 (± 0.0102)	99.33
XGBoost	0.9573 (± 0.0003)	0.7945 (± 0.0083)	33.49
Gradient Boost	0.9579 (± 0.0000)	0.8084 (± 0.0096)	132.25
LightGBM	0.9578 (± 0.0001)	0.8066 (± 0.0090)	5.24
ADABOOST	0.9578 (± 0.0002)	0.8038 (± 0.0096)	42.40
RANDOM FOREST	0.9533 (± 0.0005)	0.7349 (± 0.0104)	91.62
LOGISTIC REGRESSION	0.9579 (± 0.0001)	0.8046 (± 0.0093)	2.11
MLP CLASSIFIER	0.9565 (± 0.0005)	0.7515 (± 0.0089)	917.84
Catboost	0.9571 (± 0.0004)	0.7947 (± 0.0078)	160.13

accuracy for diabetes and stroke risk assessment which can be inferred from Figs. 6, 7, 8, 9 and 10.

The integration of user interaction and personalized recommendations enhances the models’ practical applicability, transforming them into powerful tools for proactive health management. The interpretation of risk scores and the subsequent justification of personalized recommendations bridge the gap between predictive analytics and actionable insights, empowering individuals to make informed choices that align with their health goals. In conclusion, this research offers valuable contributions to the field of predictive healthcare modelling, providing evidence of the potential for machine learning algorithms to revolutionize disease risk assessment and prevention [19].

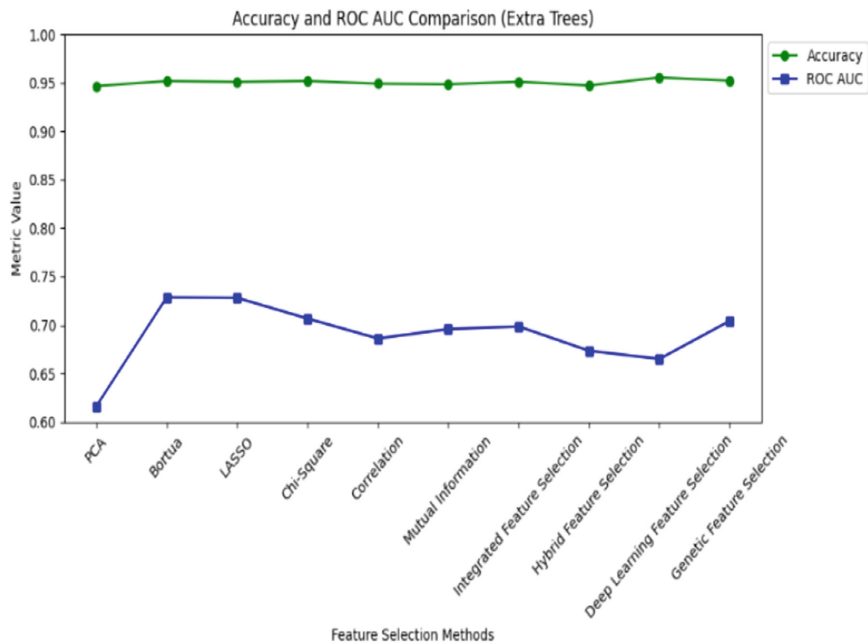


Fig. 6. Extra trees classifier (stroke)

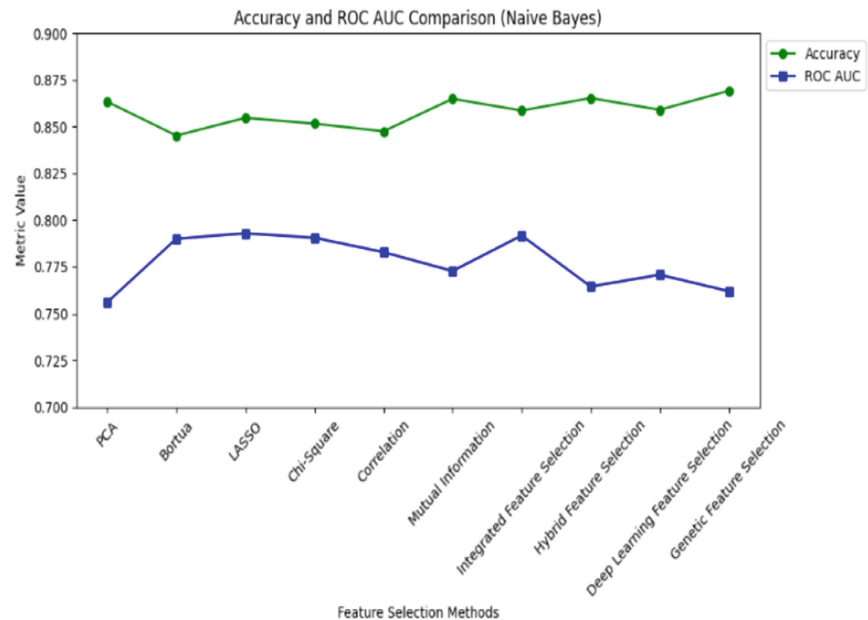


Fig. 7. Naive Bayes' (stroke)

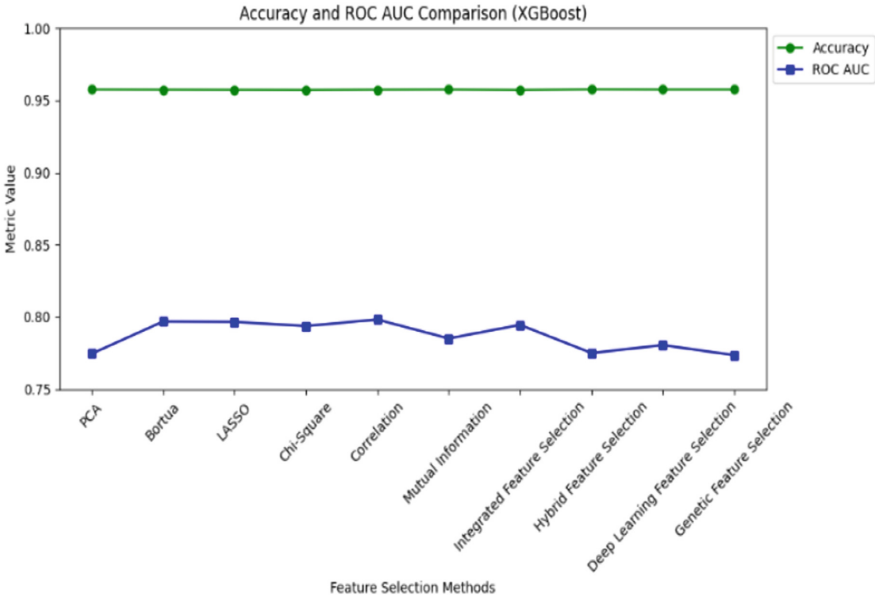


Fig. 8. XG Boost (stroke)

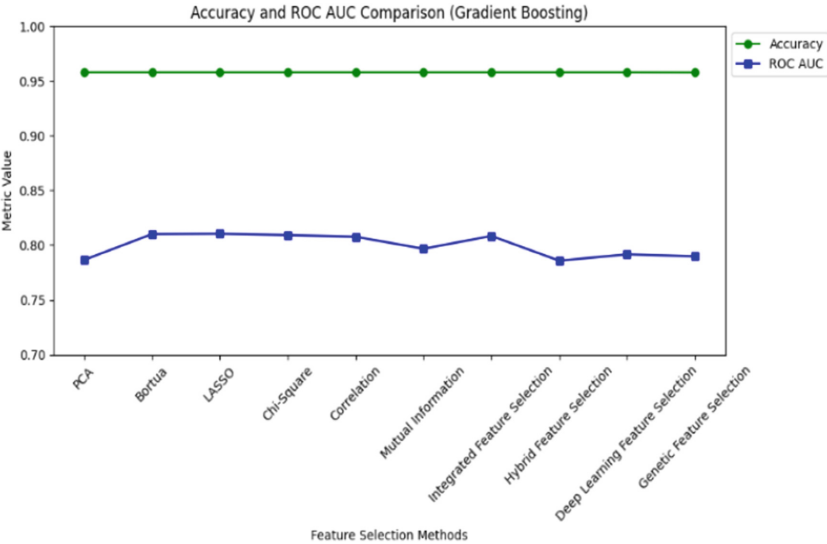


Fig. 9. Gradient Boosting (Stroke)

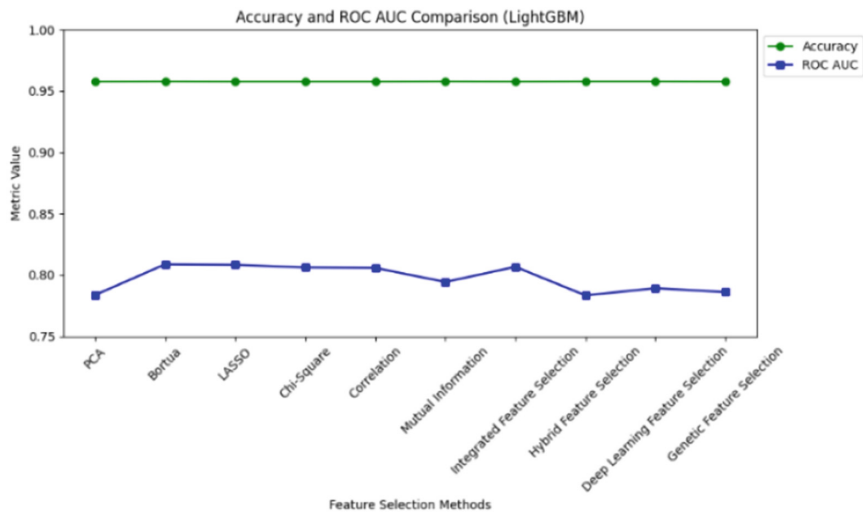


Fig. 10. LightGBM Classifier (stroke)

4 Discussion

The outcomes of our diabetes and stroke risk prediction models offer profound insights into the realm of healthcare. The accuracy and practicality of these models echo existing research on health risk assessment, reaffirming the power of cutting-edge predictive tools. Our models, finely tuned using Genetic Feature Selection (for diabetes) and Integrated Feature Selection (for stroke), demonstrate high accuracy. Genetic Feature Selection and Integrated Feature Selection emerge as game-changers, elevating predictive accuracy and guiding healthcare providers to the most pertinent risk factors. Our findings empower proactive healthcare strategies, empowering individuals to make informed choices for their well-being. In closing, our research unlocks the potential of predictive healthcare.

5 Limitations and Future Scope

The model’s performance also hinges on the diversity of the training data, which may limit accuracy if it does not encompass various demographics. User bias introduces complexity; Ethical concerns around data privacy, consent, and user comprehension add layers of complexity.

The future holds promise for predictive healthcare. Longitudinal data integration, capturing health changes over time, can enrich prediction accuracy. Embracing external data sources such as wearable and electronic health records could further enhance model inputs. By pursuing these paths, predictive models can evolve into more accurate and user-centric tools, offering actionable insights for improved health management.

6 Conclusion

In the pursuit of precise and personalized health risk prediction, this research journey ventured into the realm of predictive models for diabetes and stroke. The integration of Genetic Feature Selection (GFS) and Integrated Feature Selection (IFS) for diabetes and stroke risk prediction proved pivotal in enhancing model accuracy and comprehensibility. Notably, the LightGBM-based models exhibited impressive predictive performance, reflecting a well-balanced approach between model intricacy, data integrity, and feature relevance. The user-friendly interface facilitated seamless data input, while the models rapidly computed risk scores. The robust performance of the predictive models substantiates the feasibility of integrating machine learning algorithms into healthcare risk prediction, heralding possibilities for their incorporation into medical practices and individual health management. As with any research expedition, certain limitations came to light. Elevating user interfaces, infusing longitudinal and external data, and forging collaborations with healthcare experts can refine model precision, user engagement, and ethical dimensions. In conclusion, the reverberations of this research resonate across industries, society, and the scientific community.

References

1. Agliata, A., Giordano, D., Bardozzo, F., Bottiglieri, S., Facchiano, A., Tagliaferri, R.: Machine learning as a support for the diagnosis of Type 2 diabetes. *Int. J. Mol. Sci.* **24**(7), 6775 (2023). <https://doi.org/10.3390/ijms24076775>
2. Akkaya, B., Sener, E., Gursu, C.: A comparative study of heart disease prediction using machine learning techniques. In: 2022 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA) (2022). <https://doi.org/10.1109/hora55278.2022.9799978>
3. Ali, M.: Feature Selection Algorithms for Machine Learning - Writers Byte. Writers Byte, 26 June 2022. <https://writersbyte.com/featured-post/feature-selection-algorithms-for-machine-learning/>
4. Biswal, A.: What is a Chi-Square Test? Formula, Examples & Application. Simplilearn.com (2023). <https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test>
5. Butt, U.M., Letchmunan, S., Ali, M., Hassan, F.H., Baqir, A., Sherazi, H.H.R.: Machine learning based diabetes classification and prediction for healthcare applications. *J. Healthc. Eng.* **2021**, 1–17 (2021). <https://doi.org/10.1155/2021/9930985>
6. Center of Disease Control. (n.d.). Centers for Disease Control and Prevention (2018). Physical Activity and Health. <https://www.cdc.gov/physicalactivity/basics/pa-health/index.html>
7. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. *Comput. Electr. Eng.* **40**(1), 16–28 (2014). <https://doi.org/10.1016/j.compeleceng.2013.11.024>
8. Chang, V., Ganatra, M., Hall, K., Golightly, L., Xu, Q.: An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. *Healthc. Analytics* **2**, 100118 (2022). <https://doi.org/10.1016/j.health.2022.100118>
9. Clark, C.R., et al.: Predicting self-rated health across the life course: health equity insights from machine learning models. *J. Gen. Intern. Med.* **36**(5), 1181–1188 (2021). <https://doi.org/10.1007/s11606-020-06438-1>
10. Guo, L.: Under the background of healthy China: regulating the analysis of hybrid machine learning in sports activities to control chronic diseases. *Measurement* **164**, 107847 (2020). <https://doi.org/10.1016/j.measurement.2020.107847>

11. In supervised learning, why is it bad to have correlated features? (n.d.). Data Science Stack Exchange. <https://datascience.stackexchange.com/questions/24452/in-supervised-learning-why-is-it-bad-to-have-correlated-features>
12. Khan, M.M., Roberson, S., Reid, K., Jordan, M., Odoi, A.: Prevalence and predictors of stroke among individuals with prediabetes and diabetes in Florida. *BMC Public Health* **22**(1) (2022). <https://doi.org/10.1186/s12889-022-12666-3>
13. Khanam, J.J., Foo, S.Y.: A comparison of machine learning algorithms for diabetes prediction. *ICT Express* **7**(4), 432–439 (2021). <https://doi.org/10.1016/j.ict.2021.02.004>
14. Lee, I., Shiroma, E.J., Lobelo, F., Puska, P., Blair, S.N., Katzmarzyk, P.T.: Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *Lancet* **380**(9838), 219–229 (2012). [https://doi.org/10.1016/s0140-6736\(12\)61031-9](https://doi.org/10.1016/s0140-6736(12)61031-9)
15. Matsumoto, N., et al.: Prognostic value of myocardial perfusion single-photon emission computed tomography for the prediction of future cardiac events in a Japanese population a middle-term follow-up study. *Circ. J.* **71**(10), 1580–1585 (2007). <https://doi.org/10.1253/circj.71.1580>
16. American College of Sports Medicine. *ACSM's Guidelines for Exercise Testing and Prescription*, 10th edn. Lippincott Williams & Wilkins, Philadelphia (2013)
17. Nagarajan, S.M., Muthukumaran, V., Murugesan, R., Joseph, R.B., Munirathanam, M.: Feature selection model for healthcare analysis and classification using classifier ensemble technique. *Int. J. Syst. Assur. Eng. Manag.* (2021). <https://doi.org/10.1007/s13198-021-01126-7>
18. Nomura, A., Noguchi, M., Kometani, M., Furukawa, K., Yoneda, T.: Artificial intelligence in current diabetes management and prediction. *Curr. Diab. Rep.* **21**(12) (2021a). <https://doi.org/10.1007/s11892-021-01423-2>
19. Standards of Medical Care in Diabetes—2019 abridged for primary care providers. *Clin. Diab.* **37**(1), 11–34 (2019). <https://doi.org/10.2337/cd18-0105>
20. Tasin, I., Nabil, T.U., Islam, S., Khan, R.: Diabetes prediction using machine learning and explainable AI techniques. *Healthc. Technol. Lett.* **10**(1–2), 1–10 (2022). <https://doi.org/10.1049/htl2.12039>

Lecture Notes in Networks and Systems 1019

Kohei Arai *Editor*

Intelligent Computing

Proceedings of the 2024 Computing
Conference, Volume 4



Springer

Series Editor

Janusz Kacprzyk , *Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland*

Advisory Editors

Fernando Gomide, *Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil*

Okyay Kaynak, *Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Türkiye*

Derong Liu, *Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA*

Institute of Automation, Chinese Academy of Sciences, Beijing, USA

Witold Pedrycz, *Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada*

Systems Research Institute, Polish Academy of Sciences, Warsaw, Canada

Marios M. Polycarpou, *Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus*

Imre J. Rudas, *Óbuda University, Budapest, Hungary*

Jun Wang, *Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong*

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the worldwide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Kohei Arai
Editor

Intelligent Computing

Proceedings of the 2024 Computing
Conference, Volume 4

Editor
Kohei Arai
Faculty of Science and Engineering
Saga University
Saga, Japan

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-62272-4 ISBN 978-3-031-62273-1 (eBook)
<https://doi.org/10.1007/978-3-031-62273-1>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

It is with great pleasure that we introduce the proceedings of the Computing Conference 2024, held on July 11 and 12, 2024. This conference served as a platform for researchers and professionals from around the globe to convene and exchange ideas at the forefront of computing and its diverse applications. The enthusiasm and dedication displayed by participants underscored the significance of this event in fostering collaboration and advancing the field.

We received an overwhelming total of 457 contributions from esteemed scholars and practitioners. These submissions underwent a rigorous double peer-review process, facilitated by experts in their respective domains. After careful evaluation and deliberation, a total of 165 papers were selected for publication in these proceedings.

The diverse array of topics covered in these papers reflects the breadth and depth of contemporary computing research, spanning areas such as artificial intelligence, machine learning, cybersecurity, data science, and beyond. Each paper represents a valuable contribution to the collective knowledge base of the computing community, offering insights, innovations, and solutions to pressing challenges.

We extend our heartfelt gratitude to all authors, reviewers, organizers, and attendees whose efforts and contributions made this conference a resounding success. It is our hope that the insights shared and connections forged during this event will continue to inspire and propel advancements in computing for years to come.

Regards,
Kohei Arai

Contents

Model-Free Adaptive Fault-Tolerant Control for Offshore Wind Turbines	1
<i>Yat Ping Lam and Silvio Simani</i>	
Digital Inclusion and Longevity: Evaluating the Usability of Emerging Technologies for the Elderly	14
<i>Marily Dilamar da Silva, Edesio Marcos Slomp, Maria Teresa Silva Santos, Laís de Oliveira Dalle Mule, Douglas Ropelato, Aline Wrege Vasconcelos, and Manuella Bastos Silva</i>	
Inter-qubit Correlation of Readout Noise in Near-Term Quantum Devices	33
<i>Jonathan Miller, Bo Yang, Elham Kashefi, and Debasis Sadhukhan</i>	
Complexity Classification of Complex-Weighted Counting Acyclic Constraint Satisfaction Problems	45
<i>Tomoyuki Yamakami</i>	
Kernelized Extension for Multiple-Expert-Learning Classifiers Using Support Vector Machines	66
<i>Mohammed Bajja, Yahya Aalaila, Ana C. Umaquinga-Criollo, Lorena Guachi-Guachi, and Diego H. Peluffo-Ordóñez</i>	
Cost Optimisation in Offshore Wind Through Procurement Data Analytics	80
<i>Quadri Shittu and Chinedu Pascal Ezenkwu</i>	
Data Analysis, Socioeconomic Development and Communication of the Sugar Industry	99
<i>Sonia Tigua-Moreira, Edison Cruz-Navarrete, Lilian Molina, and Arturo Clery</i>	
Exploring the Concept of Economic Resilience Pre- and Post-2020: A Literature Review Using Text Mining and Sentiment Analysis	105
<i>Andreea Pernici and Stelian Stancu</i>	
Comprehensive Feature Selection Methods for Predicting Diabetes and Stroke Risk	128
<i>Karthik Balaji and Alexander Iliev Iliev</i>	
Prediction of Load Time in Web Services for Different Agent Locations and Browsers for Mobile Devices	146
<i>Jolanta Wrzuszczak-Noga</i>	

A Survey on the Usage of Centrality Measures in Error Prediction	157
<i>Péter Marjai and Attila Kiss</i>	
Towards Blended Learning: Students' Perceptions and Experiences of Mobile Learning in South Africa	168
<i>L. Cilliers and V. Funda</i>	
Designing an Automated Machine Learning Approach for Transformer Architecture in Education and Non-STEM Research Settings	182
<i>Luke Gassmann and Jimmy Campbell</i>	
Introducing a High School Student to Systems Programming Via Bare Machine Computing	201
<i>D. N. Naraharisetti, R. K. Karne, J. Weymouth, and A. L. Wijesinha</i>	
Accessibility of Online Educational and Scientific Resources	214
<i>Todor Todorov</i>	
A Combined Competency-Based and Design-Based Research (DBR) Approach to Curriculum Development: Implications for Information Technology (IT) Programs	223
<i>Keshnee Padayachee and Grant Royd Howard</i>	
Making Learning Effective on Smart Mobile Online Educational Games	241
<i>Bah Tee Eng and Insu Song</i>	
Students' Perceptions of ChatGPT in Education: A Rapid Systematic Literature Review	258
<i>Sibusisiwe Dube, Sinokubekezela Dube, Belinda Mutunhu Ndlovu, Kudakwashe Maguraushe, Lario Malungana, Fungai Jacqueline Kiwa, and Martin Muduva</i>	
CollaVRation: An Immersive Virtual Environment for Collaborative Software Development	280
<i>Miroslav Kozma, Juraj Vincúr, and Peter Kapec</i>	
Affect Analysis: A Literature Survey on Student-Specific and General Users' Affect Analysis	299
<i>Christine Asaju and Hima Vadapalli</i>	
Characterization of ReRAM Arrays Operating in the Pre-formed Range to Design Reliable PUFs	328
<i>Taylor Wilson, Saloni Jain, Jack Garrard, Bertrand Cambou, and Ian Burke</i>	

Introduction of Steganography to Improve the Security of Transmitted Data	351
<i>Olga Manankova, Mubarak Yakubova, and Bakhodyr Yakubov</i>	
Investigating Privacy and Security of LinkedIn Via Forensics Analysis	365
<i>Ahmad Ghafarian and Flavio Sosa</i>	
Decentralized Privacy-Preserving Solution Through Blockchain Smart Contracts and Spatial Cloaking for Ride Sharing Application	377
<i>M. Fadhil, Parman Sukarno, and Aulia Arif Wardana</i>	
An Anomaly Detection Framework for IIoT-Based Smart Farming Systems	396
<i>Muaaz Noor, Siphesihle Sithungu, and Khutso Lebea</i>	
Examining the Privacy and Security of Mobile Refugee Applications	410
<i>Marc Zirillo, Shannon Aviza, Marlon Cruz, Halie Lee, Grace LaMalva, Amanda Galvez, Denise Dragos, and Suzanna Schmeelk</i>	
Sybil Attack Detection in VANETs Using CatBoost Classifier	428
<i>EL Mehdi Fdil and EL Mamoun Souidi</i>	
Enhancing True Random Number Generation in MRAM Devices Through Response Adjustment	438
<i>Saloni Jain, Manuel Aguilar Rios, and Bertrand Cambou</i>	
DeepPayAuth: User Authentication in Mobile Payments Using Smartwatch Motion Sensors	455
<i>Richard Matovu and Ikechukwu Ohu</i>	
Is It Time for a Borderless Real Estate Platform?	473
<i>Walter Okoth Odhiambo</i>	
Vehicular Connectivity Analysis Using Enhanced Quality Slotted ALOHA (EQS-ALOHA)	484
<i>Mahmoud Zaki Iskandarani</i>	
Leveraging AI Language Models for Designing Contextually Responsive Built Environments	510
<i>Sam Sabah, Mostafa Alani, and Bilal Muiassar M. Salih</i>	
CRISP-DM User Mobility Determined IoT Placement Within a Real-World Smart Building for Resource Efficient Fog Computing Environments	520
<i>Kelvin N. Lawal, Titus K. Olaniyi, and Ryan M. Gibson</i>	
Computing VUCA Way	543
<i>Shuichi Fukuda</i>	

High Dimensional Computing Approach to Detection and Learning
Gesture Biometrics 551
 Eric Liu, William Casey, and Anthony Melaragno

Application of Speech Signal for Early Screening of Children with Autism
Spectrum Disorder 566
 Nguyen Cong-Phuong

Author Index 577