UNIVERSITY "ST. KLIMENT OHRIDSKI" - BITOLA
FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGIES - BITOLA
REPUBLIC OF NORTH MACEDONIA

Proceedings of the 13th International Conference on Applied Internet and Information Technologies

AIIT 2023

13 October, 2023, Bitola, Republic of North Macedonia

AIIT
International Conference

UNIVERSITY "ST. KLIMENT OHRIDSKI" - BITOLA
FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGIES - BITOLA
REPUBLIC OF NORTH MACEDONIA

13TH INTERNATIONAL CONFERENCE ON APPLIED INTERNET AND INFORMATION TECHNOLOGIES

AIIT 2023
PROCEEDINGS

International Conference

Bitola, 2023

Proceedings publisher and organizer of the conference:
University “St. Kliment Ohridski” - Bitola, Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia

For publisher:
Blagoj Ristevski, PhD
Faculty of Information and Communication Technologies - Bitola
University “St. Kliment Ohridski” - Bitola, North Macedonia

Proceedings editors:
Kostandina Veljanovska, PhD
Eleonora Brtka, PhD

Conference Chairmans:
Kostandina Veljanovska, University “St. Kliment Ohridski”, Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia - chair
Eleonora Brtka, University of Novi Sad, Technical Faculty “Mihajlo Pupin”, Zrenjanin, Serbia – co-chair

Technical preparation of the proceedings:

Andrijana Bocevska, PhD

Kostandina Veljanovska, PhD

Cover design:

Andrijana Bocevska, PhD

e-Proceedings
ISBN 978-608-5003-03-7

CIP - Каталогизација во публикација
Национална и универзитетска библиотека "Св. Климент Охридски", Скопје

004-049.8(062)

INTERNATIONAL conference on applied internet and information technologies AIIT 2023 (13 ; 2023 ; Bitola, Republic of North Macedonia)
Proceedings / 13th International conference on applied internet and information technologies AIIT 2023, 13 October, 2023, Bitola, Republic of North Macedonia ; [editors Kostandina Veljanovska, Eleonora Brtka]. - Bitola : University "St. Kliment Ohridski", Bitola Faculty of information and communication technologies, 2023. - 411 стр. : илустр. ; 30 см
Библиографија кон трудовите
ISBN 978-608-5003-03-7

а) Информатичка технологија -- Примена -- Собири
COBISS.MK-ID 62559493

Introduction

The International conference on Applied Internet and Information Technologies is a traditional meeting held every year, that sprouts out of collaboration between the University of Novi Sad, Technical Faculty “Mihajlo Pupin”, Zrenjanin, Serbia and the University “St. Kliment Ohridski”, Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia. The XIII AIIT2023 was held in Bitola, Macedonia on which besides the participants from Serbia and Macedonia there were researchers from Croatia, Bosnia and Herzegovina, Hungary, Finland, Russia, Turkey, Egypt, India and Australia whose contribution was either as authors or as reviewers of the papers.

At the Conference were presented innovative findings in the field of information systems, communications and computer networks, software engineering and applications, data science and big data technologies, artificial intelligence, intelligent systems, business intelligence and IT support to decision-making, data and system security, distributed systems, Internet of Things and smart systems, embedded systems, computer graphics, IT management, e-commerce, e-government, e-education, Internet marketing, and IT practice and experience.

The Conference chairs would like to express gratitude to the authors for their contributions and to express special gratitude to the reviewers for their tremendous work done for selecting the papers with their valuable comments and suggestions that contributed to improve the quality of the papers. Out of more than 60 submitted papers, 51 were selected, presented at the Conference and are published in this proceedings.

The work during the conference was organized in nine sessions: plenary session, five in-person oral sessions, one video session and two poster sessions. During the conference, a round table with participants from academic organizations and IT industry was successfully organized. The theme of the discussions at the round table was "Strengthening the capacities of Faculty of ICT for the realization of strategic cooperation with companies from the IT industry".

AIIT 2023 was very successful conference with fruitful exchange of experiences among the participants reviving the hope of further strengthening a friendly environment after the pandemic crisis. We hope that we will continue with the contribution to the further deepening the development of Internet and information technologies research.

Conference chairs:

Kostandina Veljanovska, University “St. Kliment Ohridski”, Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia (chair)
Eleonora Brtka, University of Novi Sad, Technical Faculty “Mihajlo Pupin”, Zrenjanin, Serbia (co-chair)

Conference Chairs

Kostandina Veljanovska, University “St. Kliment Ohridski”, Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia

Kostandina Veljanovska, Ph.D. finished BSc in Computer Science at the University "Sts. Kiril i Metodi", Skopje. Her first MASc in Applied Engineering she received at the University of Toronto, Toronto, Canada. Her second MSc and also her PhD in Technical Sciences she received at the University “St. Kliment Ohridski” - Bitola, R. Macedonia. Her postdoctoral studies in Artificial Intelligence she attended at the Laboratory of Informatics, Robotics and Microelectronics at the University of Montpellier, Montpellier, France. She worked as a Research assistant at the Faculty of Applied Science, University of Toronto, Canada. She also, worked as a researcher in research team for Constraints, Learning and Agents at LIRMM, University of Montpellier. Since 2008, she works as a Full Professor in Information Systems and Networks, Artificial Intelligence and Systems and Data Processing at the Faculty of Information and Communication Technologies, University “St. Kliment Ohridski” - Bitola, Rebublic of North Macedonia. Her research work is focused on artificial intelligence, machine learning techniques and intelligent systems. She has published numerous scientific papers in the area of interest, as well as several monographic items. She is a reviewing referee for well-known publishing house, journals with significant impact factor in science and also, member of editorial board of several international conferences.

Eleonora Brtka, University of Novi Sad, Technical Faculty “Mihajlo Pupin”, Zrenjanin, Serbia

Eleonora Brtka, Ph.D. is an associate professor at the Information Technology department at the University of Novi Sad, Technical Faculty „Mihajlo Pupin“, Zrenjanin. She has received her PhD in Information technology in 2015. Current research interests include: data science techniques and methods, programming languages, education, artificial intelligence and intelligent agent technologies. She is the author or co-author of several articles published in international journals and in the proceedings of international conferences. She participated in several national funded projects.

Organizing Committee

Chairs

Kostandina Veljanovska (chair), Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia
Eleonora Brtka, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia (co-chair)
Zoran Kotevski, Faculty of Information and Communication Technologies - Bitola, Republic of North Macedonia (co-chair)

Members

Željko Stojanov, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Blagoj Ristevski, Faculty of Information and Communication Technologies - Bitola, North Macedonia
Višnja Ognjenović, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Dalibor Dobrilovic, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Monika Markovska, Faculty of Information and Communication Technologies - Bitola, North Macedonia
Dragica Radosav, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Biljana Radulovic, Technical Faculty "Mihajlo Pupin" Zrenjanin, Serbia
Božidar Milenkovski, Faculty of Information and Communication Technologies - Bitola, North Macedonia
Ivana Berković, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Ljubica Kazi, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Snežana Savoska, Faculty of Information and Communication Technologies - Bitola, North Macedonia
Vladimir Brtka, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Zoltan Kazi, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Siniša Mihajlović, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Velibor Premčevski, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Nikola Rendevski , Faculty of Information and Communication Technologies - Bitola, North Macedonia
Aleksandra Stojkov, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Maja Gaborov, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Andrijana Bocevska, Faculty of Information and Communication Technologies - Bitola, North Macedonia
Milica Mazalica, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Igor Vecštejn, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Tome Dimovski, Faculty of Information and Communication Technologies - Bitola, North Macedonia
Marko Blažić, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Vuk Amižić, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Natasa Blazeska-Tabakovska, Faculty of Information and Communication Technologies - Bitola, North Macedonia
Jovana Borovina, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Dalibor Šeljmeši, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Mimoza Bogdanoska-Jovanovska, Faculty of Information and Communication Technologies - Bitola, North Macedonia
Vladimir Šinik, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Nadežda Ljubojev, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Marina Blažeković Toshevski, Faculty of Information and Communication Technologies - Bitola, North Macedonia
Vladimir Karuović, Technical Faculty "Mihajlo Pupin", Zrenjanin, Serbia
Evgeny Cherkashin, Institute of System Dynamic and Control Theory SB RAS, Russia
Anastasia Popova, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Russia
Filip Tsvetanov, South-west University "Neophyte Rilsky", Faculty of Engineering, Blagoevgrad, Bulgaria

Program Committee

Ohridski" - Bitola, North Macedonia
Tome Dimovski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, North Macedonia
Marina Blažeković Toševski, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, North Macedonia
Lela Ivanovska, Faculty of Information and Communication Technologies, University "St. Kliment Ohridski" - Bitola, North Macedonia
Ilija Hristoski, Faculty of Economics - Prilep, North Macedonia
Elena Vlahu-Gjorgievska, Univerisity of Wollongong, Australia
Ming Chen, Zhejiang University, China
Alexander Feoktistov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Alexander Yurin, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Igor Bychkov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Andrey Gachenko, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia
Andrey Mikhailov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia
Anastasia Popova, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences. Irkutsk, Russia
Alexey Daneev, Irkutsk State Transport University, Irkutsk, Russia
Denis Sidorov, Melentiev Energy Systems Institute of Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Viacheslav Paramonov, Matrosov Institute for System Dynamics and Control Theory of the Siberian Branch of the Russian Academy of Sciences, Irkutsk, Russia
Andrey Dorofeev, Institute of High Technologies, Irkutsk National Research Technical University, Irkutsk, Russia
Gogolák László, Subotica Tech - College of Applied Sciences, Subotica, Serbia
Zlatko Čović, Subotica Tech - College of Applied Sciences, Department of Informatics, Subotica, Serbia
Zora Konjović, University Singidunum, Centar Novi Sad, Serbia
Siniša Nešković, Faculty of organizational sciences, University of Belgrade, Serbia
Nataša Gospić, Faculty of transport and traffic engineering, Belgrade, Serbia
Branko Markoski, Faculty of technical Sciences, Novi Sad, Serbia
Željen Trpovski, Faculty of technical Sciences, Novi Sad, Serbia
Branimir Đorđević, Megatrend University, Belgrade, Serbia
Slobodan Jovanović, Faculty of Information Technology, Belgrade, Serbia
Željko Eremić, College of Technical Sciences - Zrenjanin, Serbia
Rajnai Zoltán, Obuda University, Budapest, Hungary
Tünde Anna Kovács, PhD, Óbuda University, Hungary
Zoltán Nyikes, PhD, Milton Friedman University, Hungary
Mirjana Pejic Bach, University of Zagreb, Croatia
Androklis Mavridis, Aristotel University of Thessaloniki, Greece
Madhusudan Bhatt, R.D. National College, University of Mumbai, India
Amar Kansara, Parth Systems LTD, Navsari, Gujarat, India
Narendra Chotaliya, H. & H.B. Kotak Institute of Science, Rajkot, Gujarat, India
Zeljko Jungic, ETF, University of Banja Luka, Bosnia and Herzegovina
Saso Tamazic, Univerisity of Ljubljana, Slovenia
Marijana Brtka, Centro de Matemática, Computação e Cognição, Universidade Federal do ABC, São Paulo, Brazil
Zoran Cosic, Statheros, Split, Croatia
Istvan Matijevics, Institute of Informatics, University of Szeged, Hungary
Slobodan Lubura, Faculty of electrical engineering, University of East Sarajevo, Bosnia and Herzegovina
Edit Boral, ASA College, New York, NY, USA

C O N T E N T

Comparative Analysis of ML Algorithms for Breast Cancer Detection

Daniela Slavkovska[1], Blagoj Ristevski,[2] Anita Petreska[2]

[1]*UTMS, Skopje Faculty of Computer Science, Republic of North Macedonia*
[2]*Faculty of Information and Communication Technologies - Bitola, University "St. Kliment Ohridski"- Bitola, Republic of North Macedonia*

dslavkovska@yahoo.com, blagoj.ristevski@uklo.edu.mk, petreska.anita@uklo.edu.mk

Abstract:
Artificial intelligence and machine learning algorithms with their advanced predictive and diagnostic techniques help healthcare providers make the right decisions in the process of disease prevention and early diagnosis. Breast cancer as a disease of modern dynamic living is gaining momentum. Its early detection is crucial to increase the chances of survival through better treatment options.
Exploratory data analysis (EDA) as a key step in data analysis involves systematic examination and visualization of data to discover patterns, outliers and dependencies, enabling hypothesis generation and making correct decisions. As powerful algorithms applied for classification, Logistic regression is used, K-nearest neighbors (KNN), Naïve Bayes, Decision Tree, Random Forest, XGBoost Support, Deep Learning and NeNetwork. In our study, a comparative analysis of the most commonly used Machine Learning (ML) algorithms has been done by evaluating various metrics such as accuracy, F-measure, confusion matrix and specificity.
Limitations of machine learning algorithms often include issues such as overhead, high computational requirements, and data quality. Challenges may arise from the need for large labeled datasets, algorithmic bias, and concerns about interpretability.
Future work in machine learning should focus on developing more robust models that can generalize well to diverse data, improving the interpretability of complex models.
Keywords:
Machine Learning, Breast Cancer, Artificial intelligence, Medical big data, Healthcare, Data mining, Exploratory Data Analysis

1. Introduction

In recent years, healthcare has been facing numerous challenges, both from the increased accuracy of diagnosis and therapy and the need for an efficient way of managing large amounts of health data[1]. The ability to extract useful knowledge hidden in large amounts of data and act on that knowledge is becoming increasingly challenging. Cancer, especially breast cancer [2][7], is one of the most dangerous diseases in the world, claiming the lives of countless women every year. Machine learning algorithms [8] play a major role in the early detection and diagnosis of cancer. Section 2 reviews the role of machine learning and Data Mining in healthcare [6][15], and Section 3 presents the most commonly used ML algorithms in healthcare [3]. Part 4 presents the application of machine learning algorithms in order to discover patterns and anomalies that are crucial in the detection and diagnosis of breast carcinoma, the used methodology, data collection, data preprocessing, Exploratory Data Analysis [9], evaluation metrics and comparison are presented of the results. In the fifth part, the limitations and challenges faced by ML algorithms are presented, and in the last, sixth part, the conclusions reached and future challenges of the ML algorithms are presented.

2. The role of machine learning in healthcare

Artificial intelligence [5][20] has a great impact on medical research that has direct application in real medical applications. Given the exponential growth of data, traditional manual data analysis is not enough. Methods for efficient computational analysis such as technologies developed in the field of data mining are necessary to discover insights from Big Data. Data Mining (DM) enables the discovery of knowledge in databases that includes understanding the domain, preprocessing the data, extracting regularities hidden in the data in order to create patterns or models. Numerous Big Data applications have been developed with innovative methods for improving the quality of health care in the direction of prevention or early detection of diseases.

Machine learning as one of the key components of artificial intelligence is used to train algorithms and models that can analyze data, extract patterns and make predictions based on experience. Machine learning leverages AI's ability to learn from data and improve its performance with experience, making AI more intelligent and adaptable to different tasks and environments.

3. Most commonly used algorithms of machine learning in healthcare

Various machine learning algorithms are used in the healthcare industry covering various aspects of medicine and healthcare.

Logistic regression [8] is a simple and interpretable algorithm that is often used for binary classification tasks, such as breast cancer prediction. It models the probability that a given case belongs to a certain class. Characteristics related to patient demographics, genetic factors, and medical history can be used as input variables.

Support vector machine (SVM) [6][16]is a powerful algorithm for both binary and multiclass classification that uses a hyperplane that maximizes the margin between different classes. SVM can effectively handle high-dimensional spaces, making it suitable for genetic and medical data.

Decision trees [6][12] are used for classification by recursively partitioning the data into subsets based on the most informative features. They are interpreted and can provide insight into which features are most important in predicting breast cancer. Decision trees may require pruning to prevent overgrowth.

Random forests [7][8] are an ensemble method that builds multiple decision trees and combines their predictions. They reduce the risk of overload associated with single-decision trees. Random forests can handle high-dimensional data and capture complex interactions between features.

KNN [6][8] is a non-parametric algorithm that classifies data points based on the majority class among their k-nearest neighbors. It can be used to predict breast cancer by looking at similar cases of patients. The choice of the 'k' parameter is important and should be determined through cross-validation.

The Naive Bayes [8][17] classifier works by calculating the probability of breast cancer occurrence based on characteristics such as age, family history, tumor characteristics, and medical test results. Although it simplifies the assumption of independence of features, it can still be effective in aiding breast cancer diagnosis and risk assessment.

XG Boost [8][18] is known for its speed, efficiency, and high prediction performance. The ability to handle a wide range of input features allows it to be applied to breast cancer detection by training a model containing numerous features extracted from breast cancer-related medical images, patient data, or genetic information.

Neural networks [17][19][21], especially deep learning models, can automatically learn complex patterns and relationships in data, making them suitable for tasks involving complex visual data. They have outstanding performance in image recognition tasks, including medical image analysis. Table 1 lists the advantages and disadvantages of the most commonly used ML algorithms.

3.1. Advantages and disadvantages of machine learning algorithms

An advantage of logistic regression is its simplicity and interpretability, which makes it suitable for solving problems related to linear dependence and basic modeling, and a disadvantage is its limited ability to model complex, non-linear data patterns.

Support Vector Machine is efficient for complex, high-dimensional data and can achieve high accuracy with proper kernel selection, but can be computationally expensive for large datasets.

The advantages of decision trees include their simplicity and interpretability, while their disadvantages include susceptibility to overfitting and instability due to small variations in the data.

The advantage of Case Forest is high accuracy, robustness to overload, and suitability for different data types and large data sets, and the disadvantage is reduced interpretability compared to a single decision tree and can be computationally intensive.

The K-Nearest Neighbors algorithm is simple and effective to implement for locally structured data without making assumptions about the data distribution and can be effective for classification tasks, but it can be computationally expensive, especially with large datasets and many rely on the choice of the "k" parameter, which can affect its performance making it less suitable for high-dimensional databases.

The advantage of XGBoost lies in its excellent power as an ensemble learning method, efficiency, and adaptability for complex problems, known for its high predictive accuracy and efficiency due to gradient boosting, but it can be prone to overfitting if not properly tuned and can to require more computational resources compared to simpler algorithms.

An advantage of a neural network is its ability to model complex relationships in data to deal with complex tasks with large data sets, including image and text analysis, and the disadvantage of a neural network is that it requires a large amount of data and computational resources, lacks transparency, which can hinder their interpretability, especially in critical applications where understanding the model's decision-making process is crucial.

3.2. Comparison of Machine Learning Algorithms in terms of Accuracy, Interpretability and Adaptability

The accuracy of machine learning algorithms can vary depending on the dataset and parameters, but in general, Random Forest, XGBoost, and Neural Networks tend to achieve high accuracy due to their ability to capture complex patterns, logistic regression and decision trees can tend to work well for simpler problems, in more complex scenarios they have lower accuracy. SVM and neural networks are very adaptable to different data types and complexity, while decision trees and random forests may require additional techniques to deal with overload, KNN is sensitive to the volume and dimensionality of the data. Algorithm selection should consider trade-offs between these aspects based on the specific problem and dataset.

In terms of interpreting the results of decision tree and logistic regression is relatively simple, they are generally more interpretable because they provide explicit rules and coefficients that connect the input characteristics to the predictions. Neural networks, SVM, Random Forest, KNN, and XGBoost are less interpretable due to their complex, non-linear, and ensemble-based nature, which allows meaningful insights to be extracted.

In terms of adaptability SVMs and neural networks are highly adaptable to different data types and complexity, making them suitable for a wide range of tasks. Decision trees and random forests are scalable but may require techniques to resolve overload, while KNNs can be sensitive to data volume and dimensions, affecting their adaptability to certain scenarios.

4. Most commonly used algorithms of machine learning in healthcare

In this chapter, a study is made based on medical data that may come from hospitals, research institutions, or publicly available data sets in order to predict the risk of breast cancer. In the study, machine learning methods are proposed for the purpose of breast cancer diagnosis.

4.1. Methodology

Methodology [4] for cancer assessment includes 10 stages shown in Table 1.

Table 1:
Methodology for breast cancer assessment

Phase	description
First stage	Data collection and pre-processing
Second phase	Data cleaning and pre -processing
Third phase	Normalize or standardize
Fourth phase	Feature selection and engineering
The fifth stage	Split the database
The sixth stage	Model selection and training
Seventh stage	Model evaluation
Eighth stage	Interpretation and visualization of the model
Ninth stage	Deployment and testing
The tenth stage	Ethical considerations and compliance

The first phase involves gathering relevant medical data, including patient demographics, genetic information, medical history, and diagnostic characteristics of breast cancer. The data cleaning is performed in order to deal with missing values, analysis of outliers, and data quality. To ensure uniform scaling, it is necessary to perform normalization and standardization of numerical characteristics and coding of categorical variables. Then divide the database into training, validation, and test sets for model evaluation. Unnecessary features are removed in order to reduce dimensionality. Univariate or multivariate analysis techniques are used in order to examine the correlation between variables.

The next step is the selection of appropriate machine learning algorithms. The selected models are trained based on the training data. In order to prevent overfitting of the model, the performance of the model is evaluated on the validation set. Models are evaluated using appropriate metrics for binary classification, such as accuracy, precision, recall, F1-score area under the ROC curve (AUC-ROC), confusion matrix, and cross-validation. It is good practice to visualize the results of the model to gain insight and interpretability. It requires continuous monitoring of the model and implementing mechanisms to update it when new data becomes available. We need to ensure that the use of patient data complies with ethical and legal regulations, such as HIPAA (in the United States) or GDPR (in Europe). The final phase involves documenting the entire methodology, including data sources, preprocessing steps, model selection, and evaluation results.

4.2. Data collection

In our study, a database downloaded from Kaggle **Error! Reference source not found.** was used, consisting of 33 rows containing 33 attributes of which 32 are input attributes and one target attribute that indicates the risk of breast cancer. Attributes contain demographic data and clinical parameters that encompass different aspects of an individual's health profile.

4.3. Data pre-processing

In the process of data preprocessing, categorical variables are converted into /indicator variables, missing values are checked if there are any, which is of crucial importance to either be rejected or filled with a mean or interpolated value. Table 2 presents the basic statistical parameters of the database

Table 2:
Basic statistical parameters

	count	mean	std	min	25%	50%	75%	max
Radius mean	569.000000	14.127292	3.524049	6.981000	11.700000	13.370000	15.780000	28.110000
Texture mean	569.000000	19.289649	4.301036	9.710000	16.170000	18.840000	21.800000	39.280000
Perimeter mean	569.000000	91.969033	24.298981	43.790000	75.170000	86.240000	104.100000	188.500000
Area mean	569.000000	654.889104	351.914129	143.500000	420.300000	551.100000	782.700000	2501.000000
Smoothness mean	569.000000	0.096360	0.014064	0.052630	0.086370	0.095870	0.105300	0.163400
Compactness mean	569.000000	0.104341	0.052813	0.019380	0.064920	0.092630	0.130400	0.345400
Concavity mean	569.000000	0.088799	0.079720	0.000000	0.029560	0.061540	0.130700	0.426800
Concave points mean	569.000000	0.048919	0.038803	0.000000	0.020310	0.033500	0.074000	0.201200
Symmetry mean	569.000000	0.181162	0.027414	0.106000	0.161900	0.179200	0.195700	0.304000
Fractal dimension mean	569.000000	0.062798	0.007060	0.049960	0.057700	0.061540	0.066120	0.097440
Radius se	569.000000	0.405172	0.277313	0.111500	0.232400	0.324200	0.478900	2.873000
Texture se	569.000000	1.216853	0.551648	0.360200	0.833900	1.108000	1.474000	4.885000
Perimeter se	569.000000	2.866059	2.021855	0.757000	1.606000	2.287000	3.357000	21.980000
Area se	569.000000	40.337079	45.491006	6.802000	17.850000	24.530000	45.190000	542.200000
Smoothness se	569.000000	0.007041	0.003003	0.001713	0.005169	0.006380	0.008146	0.031130
Compactness se	569.000000	0.025478	0.017908	0.002252	0.013080	0.020450	0.032450	0.135400
Concavity se	569.000000	0.031894	0.030186	0.000000	0.015090	0.025890	0.042050	0.396000
Concave points se	569.000000	0.011796	0.006170	0.000000	0.007638	0.010930	0.014710	0.052790
Symmetry se	569.000000	0.020542	0.008266	0.007882	0.015160	0.018730	0.023480	0.078950
Fractal dimension se	569.000000	0.003795	0.002646	0.000895	0.002248	0.003187	0.004558	0.029840
Radius worst	569.000000	16.269190	4.833242	7.930000	13.010000	14.970000	18.790000	36.040000
Texture worst	569.000000	25.677223	6.146258	12.020000	21.080000	25.410000	29.720000	49.540000
Perimeter worst	569.000000	107.261213	33.602542	50.410000	84.110000	97.660000	125.400000	251.200000
Area worst	569.000000	880.583128	569.356993	185.200000	515.300000	686.500000	1084.000000	4254.000000
Smoothness worst	569.000000	0.132369	0.022832	0.071170	0.116600	0.131300	0.146000	0.222600
Compactness worst	569.000000	0.254265	0.157336	0.027290	0.147200	0.211900	0.339100	1.058000
Concavity worst	569.000000	0.272188	0.208624	0.000000	0.114500	0.226700	0.382900	1.252000
Concave points worst	569.000000	0.114606	0.065732	0.000000	0.064930	0.099930	0.161400	0.291000
Symmetry worst	569.000000	0.290076	0.061867	0.156500	0.250400	0.282200	0.317900	0.663800
Fractal dimension worst	569.000000	0.083946	0.018061	0.055040	0.071460	0.080040	0.092080	0.207500

4.4. Exploratory data analysis

The EDA process [13] begins with summarizing the database, providing basic statistics and identifying data types, figuring out the size and structure of the database. A missing values analysis follows, which allows us to assess the extent of missing data and decide on appropriate techniques to be applied in order to ensure the completeness of the modeling data. Descriptive statistics and data distribution visualizations reveal the statistical properties of numerical features, making it easier to identify outliers and understand the distribution of features.

Correlation analysis reveals the relationships between features and allows us to measure multicollinearity between them. Particular attention is paid to the distribution of the target variable, understanding the prevalence of breast cancer cases and its impact on the class imbalance of the model.

During EDA, data visualization plays a key role in discovering relationships, identifying outliers, and gaining insight into potential feature engineering opportunities. The insights gained from the comprehensive EDA, feature selection, and model development steps indicate that the breast cancer prediction ML model is built on a solid understanding of the complexity of the database. Ethical considerations when handling sensitive medical data need to be addressed to address issues of privacy and bias, contributing to responsible AI applications in healthcare. Figure 1 shows a flowchart of ML algorithms for breast cancer prediction.

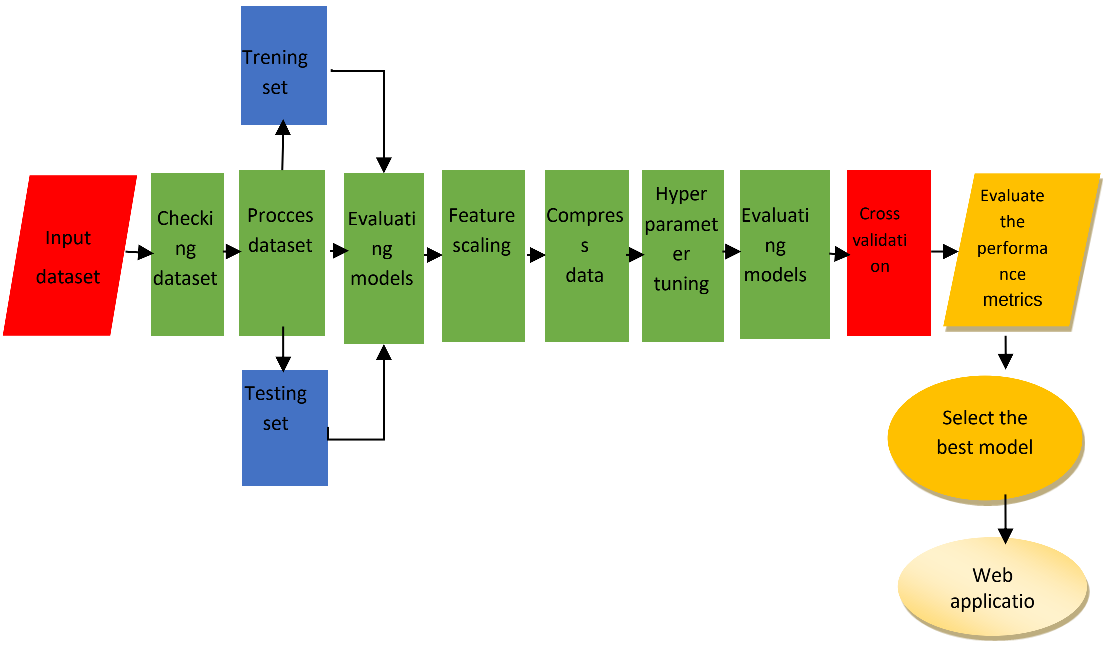

Figure 1: Flowchart of ML algorithms for breast cancer prediction

A correlation was made in order to reveal the potential relationships between the variables so that interactions and dependencies could be seen. This analysis provides valuable insights into the potential influence between variables as well as their joint influence on breast cancer risk. Figure 2 shows a correlation matrix.

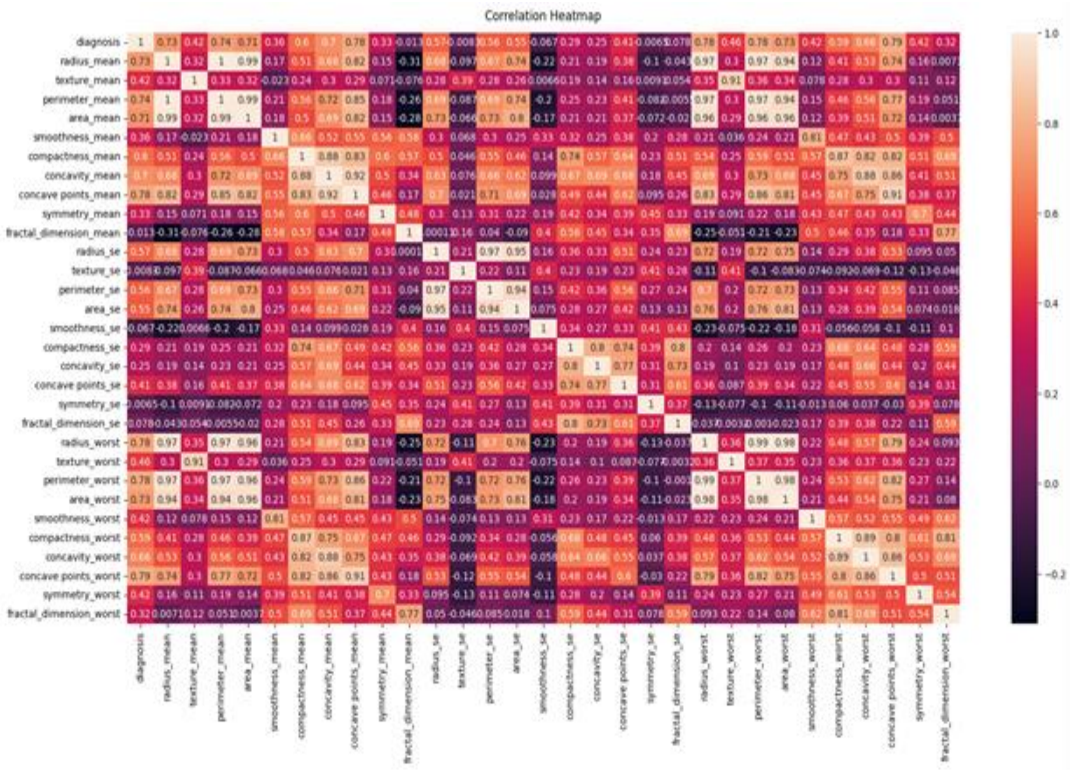

Figure 2: Correlation matrix

4.5. Evaluation Metrics

Using appropriate metrics [14] to evaluate machine learning models aims to determine the best algorithm for a specific task according to the objectives of the problem. The choice of evaluation metrics can vary depending on the machine learning model chosen. An overview of the most commonly used metrics for evaluating classification and regression tasks is provided:

Accuracy: Measures how many of the total number of instances in the test set were correctly classified as true. Accuracy = (TP + TN) / (TP + TN + FP + FN), where TP (True Positives) is the number of true positives, TN (True Negatives) is the number of true negatives, FP (False Positives) is the number of false positives, and FN (False Negatives) is the number of false negatives.

Precision: Measures how many instances classified as positive result in true positives. Accuracy = TP / (TP + FP).

Sensitivity (Sensitivity, Recall): Measures how many of the true positive instances were identified as positive. Recall = TP / (TP + FN).

F1-Score: A joint measure of precision and recall that measures a harmonic average between them. F1-score = 2 * (Precision * Recall) / (Precision + Recall).

Area under the receiver operating characteristic curve (AUC-ROC): Measures the performance of the model for different decision threshold levels. In other words, it is the probability that the model will give an earlier score to a randomly selected positive instance than to a randomly selected negative instance.

Area under the precision-recall curve (AUC-PR): Measures model performance for different decision threshold levels, but focuses on precision and recall. In other words, it is the probability that the model will give an earlier score to a randomly selected positive instance than to a randomly selected negative instance.

Confusion Matrix: A table showing the number of true and false positive and negative classifications of the model. It can be used to calculate all of the above metrics.

The choice of evaluation metric(s) should align with the specific goals of your machine learning project. It is advisable to consider multiple metrics to get a comprehensive overview of model performance.

4.6. Comparison of results

Table 3 shows the Evaluation Scores of the considered ML algorithms by evaluating their performance in terms of accuracy, Cross Validation Score, and ROC_AUC Score in order to build a model for breast cancer risk prediction.

Table 3:
Accuracy, f1 score, precision, recall, balanced accuracy in different ML models

ML model	accuracy	f1 score	precision	recall	balanced accuracy
K-Nears Neighbors	0.973684	0.963855	1.000000	0.930233	0.965116
Naïve Bayes	0.973684	0.963855	1.000000	0.930233	0.965116
Logistic Regression	0.964912	0.952381	0.975610	0.930233	0.958074
Random Forest	0.964912	0.952381	0.975610	0.930233	0.958074
XG Boost	0.964912	0.952381	0.975610	0.930233	0.958074
Neural Network	0.956140	0.941176	0.952381	0.930233	0.951032
Decision Tree	0.929825	0.909091	0.888889	0.930233	0.929905

K-Nears Neighbors and Naïve Bayes: Both models achieved the highest accuracy (97.37%) and F1 score (0.9639) among the models. They also have perfect precision (1.000) and high recall (0.9302), indicating a good balance between correctly classifying positive examples and minimizing false positives. Their balanced precision is also the highest at 0.9651, indicating good performance in both classes.

Logistic Regression, Random Forest, and XG Boost: These three models have similar performance indicators, with an accuracy of 96.49%, an F1 score of 0.9524, and a precision of 0.9756.

Their recall is also 0.9302, indicating that they are efficient at finding positive examples. The balanced accuracy is slightly lower than K-Nearest Neighbors and Naïve Bayes at 0.9581, but it is still a strong performer.

The Neural Network model has slightly lower accuracy (95.61%) and F1 score (0.9412) compared to previous models. Its accuracy and recoil are similar to previous models, both at 0.9524 and 0.9302, respectively. The balanced precision is 0.9510, which is still a good performance.

The Decision Tree model has the lowest accuracy (92.98%) and F1 score (0.9091) among the models.

Its precision (0.8889) is also lower, indicating a higher rate of false positives. However, its recall is still 0.9302, suggesting that it is effective in finding positive examples. Balanced accuracy (0.9299) is the lowest among the models, indicating slightly lower performance of both classes.

In summary, K-Nearest Neighbors and Naïve Bayes stand out as the best-performing models for this classification task, followed by logistic regression, random forest, and XG Boost. The neural network performs slightly below these models but is still a strong contender. The Decision Tree model lags behind in terms of performance. Choosing the best model should consider factors beyond these metrics, such as model complexity, interpretability, and computational requirements.

In this study, we attempted to predict breast cancer risk using machine learning methods and medical data from Kaggle. This study emphasized the importance of data quality and the application of statistical and machine analyses to identify important correlations in medical data. Evaluation of the models with different metrics provided information about their effectiveness and reliability. It should be kept in mind that this study is only one step towards improving the diagnosis of breast cancer and that continuous research and innovation is needed.

5. Limitations and challenges

Machine learning algorithms in healthcare face numerous limitations and challenges, related to the quality and availability of healthcare data, potential biases in data collection leading to biased models and health disparities, and the interpretability of models. Ethical concerns and privacy regulations present significant obstacles in managing patient data while ensuring confidentiality.

Imbalanced data, temporal aspects of health data, resource constraints, regulatory compliance, and the need to continuously adapt the model to evolving health environments are significant obstacles. The integration of machine learning into existing clinical workflows can be complex and may require changes to established practices, ensuring model validation and accountability is critical in healthcare.

Machine learning (ML) algorithms in the field of healthcare have shown significant progress, but still struggle with significant limitations and face various challenges, such as data quality, data privacy, interoperability, ethical issues, clinical validation, generalizability, model explanation, resource constraints, regulatory hurdles, lack of standardization, human oversight, barriers to adoption, cost-benefit analysis.

Addressing these challenges requires a collaborative, multidisciplinary approach involving healthcare professionals, data scientists, ethicists, policymakers, and various stakeholders to develop robust, ethical, and clinically valuable machine learning applications in the healthcare sector.

6. Conclusions and future work

Machine learning (ML) and artificial intelligence (AI) are revolutionizing healthcare by automating diagnostic processes, predicting disease risks, and personalizing treatment plans based on patient data. These technologies improve medical image analysis, support healthcare professionals in decision-making, and streamline healthcare data management. Although they offer significant benefits, their ethical use and privacy of patient data remain important considerations in their implementation.

Machine learning has made great strides in the healthcare field, offering great potential to improve patient care, diagnostics, and treatment, ensuring quality and standardization of healthcare data, which are critical to training accurate and reliable ML models. This implies a dedicated effort in data collection, cleaning, and harmonization to obtain the desired results.

Ethical dilemmas and privacy are significant aspects, requiring continued research into privacy-preserving techniques and ethical frameworks aimed at protecting patient data while enabling meaningful analysis.

Interpretability and explain ability of ML models are of critical importance, especially in the healthcare decision-making process, continuous research on interpretable AI methodologies and model explanation is needed. Mitigating bias in health data and algorithms to promote fairness, the development of unbiased models becomes necessary to ensure equitable health care outcomes.

Validation studies are necessary to confirm the safety and efficacy of ML models in clinical settings. Collaboration between researchers, clinicians, and regulatory bodies is essential.

ML models must adapt to new data and protocols in healthcare. Research devoted to adaptive and lifelong learning approaches becomes imperative.

Integrating ML into clinical workflows is a huge challenge, but it is still essential that research should concentrate on developing user interfaces and workflows that healthcare professionals can easily adopt.

Effective allocation of resources in resource-constrained healthcare facilities requires careful consideration.

In the future, priority should be given to patient-focused AI applications that enable individuals to make informed decisions about their health.

Interdisciplinary collaboration among data scientists, healthcare professionals, ethicists, policymakers, and other stakeholders remains necessary for the successful development, validation, and deployment of ML algorithms in healthcare.

The future calls for continued advances in AI and ML in the healthcare sector, with an increased focus on personalized medicine, early disease detection and improving patient outcomes. Addressing these challenges will be critical to realizing the full potential of ML in healthcare.

References:

[1] Kirola, Madhu, et al. "A comprehensive review study on: optimized data mining, machine learning and deep learning techniques for breast cancer prediction in big data context." Biomedical and Pharmacology Journal 15.1 (2022): 13-25.

[2] Thagaard, Jeppe, et al. "Pitfalls in machine learning-based assessment of tumor-infiltrating lymphocytes in breast cancer: a report of the international immuno-oncology biomarker working group." The Journal of Pathology (2023).

[3] Ozaydin B, Berner ES, Cimino JJ. Appropriate use of machine learning in healthcare. Intelligence-Based Medicine. 2021 Jan 1; 5: 100041.

[4] Gou, Qiheng, et al. "Systematic evaluation of tumor microenvironment and construction of a machine learning model to predict prognosis and immunotherapy efficacy in triple-negative breast cancer based on data mining and sequencing validation." Frontiers in Pharmacology 13 (2022): 995555.

[5] Clark, Phoebe, Jayne Kim, and Yindalon Aphinyanaphongs. "Marketing and US Food and Drug Administration Clearance of Artificial Intelligence and Machine Learning Enabled Software in and as Medical Devices: A Systematic Review." JAMA Network Open 6.7 (2023): e2321792-e2321792.

[6] Yue, Wenbin, et al. "Machine learning with applications in breast cancer diagnosis and prognosis." Designs 2.2 (2018): 13.

[7] Amrane, Meriem, et al. "Breast cancer classification using machine learning." 2018 Electric Electronics, computer science, biomedical engineering meeting (EBBT). IEEE, 2018.

[8] Asri, Hiba, et al. "Using machine learning algorithms for breast cancer risk prediction and diagnosis." Procedia Computer Science 83 (2016): 1064-1069.

[9] Shalini, M., and S. Radhika. "Machine learning techniques for prediction from various breast cancer datasets." 2020 Sixth international conference on biosignals, images, and instrumentation (ICBSII). IEEE, 2020.

[10] Christobel, Angeline, and Y. Sivaprakasam. "An empirical comparison of data mining classification methods." International Journal of Computer Information Systems 3.2 (2011): 24-28.

[11] Baughan, Natalie, Lindsay Douglas, and Maryellen L. Giger. "Past, present, and future of machine learning and artificial intelligence for breast cancer screening." Journal of Breast Imaging 4.5 (2022): 451-459.

[12] Bari Antor, Morshedul, et al. "A comparative analysis of machine learning algorithms to predict Alzheimer's disease." Journal of Healthcare Engineering 2021 (2021).

[13] Indrakumari, R., T. Poongodi, and Soumya Ranjan Jena. "Heart disease prediction using exploratory data analysis." Procedia Computer Science 173 (2020): 130-139.

[14] Li, Yixuan, and Zixuan Chen. "Performance evaluation of machine learning methods for breast cancer prediction." Appl Comput Math 7.4 (2018): 212-216.

[15] Ahmad, L. Ghasem, et al. "Using three machine learning techniques for predicting breast cancer recurrence." J Health Med Inform 4.124 (2013): 3.

[16] Sivakami, K., and Nadar Saraswathi. "Mining big data: breast cancer prediction using DT-SVM hybrid model." International Journal of Scientific Engineering and Applied Science (IJSEAS) 1.5 (2015): 418-429.

[17] Men, Kuo, et al. "Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning." Physica Medica 50 (2018): 13-19.

[18] Rane, Nikita, et al. "Breast cancer classification and prediction using machine learning." International Journal of Engineering Research and Technology 9.2 (2020): 576-580.

[19] Mahesh, T. R., et al. "Performance analysis of xgboost ensemble methods for survivability with the classification of breast cancer." Journal of Sensors 2022 (2022): 1-8.
[20] Saoud, Hajar, et al. "Using feature selection techniques to improve the accuracy of breast cancer classification." Innovations in Smart Cities Applications Edition 2: The Proceedings of the Third International Conference on Smart City Applications. Springer International Publishing, 2019.
[21] Abdulaziz, Meliboyev. "ACCURATE BREAST CANCER CLASSIFICATION BY USING ARTIFICIAL INTELLIGENCE ALGORITHMS." Journal of Universal Science Research 1.5 (2023): 1579-1586.
[22] Baraniya, S., et al. "Breast Cancer Classification and Recurrence Prediction Using Artificial Neural Networks and Machine Learning Techniques." 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT). IEEE, 2023.
[23] Petrovski, G., Savoska, S., Ristevski, B., Bocevska, A., Jolevski, I., & Blazheska-Tabakovska, N. (2022). Visual Data Analysis for EU Public Sector Data using the app MyDataApp.
[24] Ristevski, Blagoj, Snezana Savoska, and Natasha Blazheska-Tabakovska. "Opportunities for big data analytics in healthcare information systems development for decision support." (2020): 38-50.
[25] https://www.kaggle.com/datasets/nancyalaswad90/breast-cancer-dataset?select=data.csv