X-Ray Images Analytics Algorithm based on Machine Learning

VESKA GANCHEVA, IVAYLO GEORGIEV, VIOLETA TODOROVA
Technical University of Sofia,
BULGARIA

Abstract: - The rapid development of information technology has led to a huge amount of data generated by large or complex systems and devices. Applications in information technology, medicine, and many other fields generate large volumes of data that challenge analysts. Data mining analysis finds application in areas where statistical and analytical methods and the models built through them are not sufficient. The paper discusses sources of medical data, use cases, and data analysis in medicine, as well as methods and algorithms for data analysis. The purpose and objectives of the study, presented in the paper are to propose an algorithm for processing X-Ray images based on tools and techniques from the field of machine learning. The preprocessing phase is concerned with image transformation, feature extraction, and the selection of training and testing datasets. Preprocessing data enables the processing of data that would not otherwise be appropriate by adjusting the data to the specifications established by each data retrieval procedure. Each feature is examined in the second stage to identify and classify any potential patterns. In the final stage, the most effective model to capture the pattern or behaviour of the data is chosen using a machine learning algorithm. The proposed algorithm is verified using publicly available X-Ray image datasets consisting of four classes: Normal, Lung Opacity, Pneumonia, and COVID-19. A medical image classification workflow was designed for verification. In the experimental workflow, five algorithms in the field of machine learning are determined and implemented: Logistic Regression, Naive Bayes, Random Forest, SVM, and Neural Network. In comparison to the outcomes of Random Forest, Logistic Regression, Naive Bayes, and SVM, the findings of the experimental analysis and results demonstrate that Neural Networks produce the greatest results, and these results can be taken to be the most dependable.

Key-Words: - Covid-19, data analytics, image classification, machine learning, X-Ray images.

Received: May 24, 2022. Revised: February 18, 2023. Accepted: March 26, 2023. Published: April 28, 2023.

1 Introduction

Nowadays, the growth of computer simulations has led to the accumulation of huge amounts of data. Medicine is a fundamental field highly dependent on big data technologies. This stimulates the progress of data processing technologies and methods.

Techniques for big data provide opportunities to collect large sets of biological samples and store, manage, and analyze the data, [1].

Machine learning algorithms allow the generation of additional data other than the original input data, thereby creating knowledge from data.

Rapid Learning Healthcare (RLHC) models using artificial intelligence can detect data of varying quality when compared to validated datasets. The extracted result is processed through decision support systems (DSS) to provide knowledge-based healthcare.

1.1 Sources of Medical Data

Sources of big data in healthcare include smart devices, genetic databases, government, and more (Fig. 1).

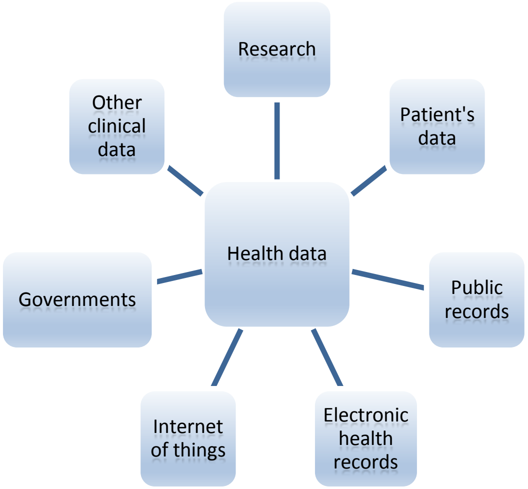


Fig. 1: Sources of data for healthcare

1) Internet of Things (IoT):

- Wearable devices
- Smartphone applications
- Medical devices and sensors

2) Electronic Medical Records/Electronic Health Records (EMR/EHR).

3) Other clinical data.

1.2 Big Data Use Cases and Data Analytics in Medicine

The insights gained from analyses of data provide medical professionals with knowledge not previously available, [2], [3]. In medicine, data is used throughout the entire healthcare cycle: medical examinations, laboratory data, patient history, and outcomes. Real-world applications demonstrate how an analytics approach can improve processes, improve patient care, and ultimately save lives.

1) Diagnostics
2) Modeling and forecasts
3) Real-time monitoring of the patient's vital signs
4) Treatment of serious diseases
5) Population health
6) Preventive care
7) Electronic Health Records (EHRs)
8) Telemedicine
9) Real-time alerts
10) Data integration with medical images

1.3 Related Works

The respiratory system is the principal organ affected by COVID-19, a viral illness. Numerous medical experts in numerous sectors continue to look for novel ways to identify and treat the disease. The radiographic lung effects of COVID-19 have received a lot of attention from researchers, [4].

Healthcare is increasingly utilizing artificial intelligence and machine learning. The analyses carried out using cutting-edge technology are improving in precision and applicability across many areas of medicine, facilitating quicker and occasionally more accurate clinical diagnosis.

It is crucial to analyze X-ray images to diagnose respiratory diseases, especially pneumonia brought on by COVID-19 infection, [5], [6], [7], [8]. Machine learning and convolutional neural networks, two disciplines of artificial intelligence that deal with image analysis, are the foundation of this type of diagnostic system. The most well-known uses of convolutional neural networks and machine learning include the detection of images, image segmentation, and image classification.

The diagnosis of respiratory illnesses and the detection of pneumonia brought on by the COVID-19 disease are two areas where image classification is directly applied. The network training process requires significant system and time resources in addition to a sufficient number of pre-classified X-ray images.

A model for automatic COVID-19 detection using raw chest X-ray images is designed, [9], to give precise diagnostics for binary classification (COVID vs. No-Findings) and multi-class classification (COVID vs. No-Findings vs. Pneumonia). For binary classes, the model provided a classification accuracy of 98.08%, and for multi-class scenarios, it produced an accuracy of 87.02%.

Five different algorithms for machine learning (such as Support Vector Machine (SVM), K-Nearest Neighbors, Random Forest, Naive Bayes Algorithm, and Decision Tree) were used for classifying X-ray images between COVID-19 and normal, [10]. Evaluation of results shows that SVM ensures the highest accuracy from 96% among the remaining four classifications (K-Nearest Neighbors и Random Forest - 92% accurate, 90% accurate on the algorithm Naive Bayes и 82% accurately on Decision Tree).

The research described in the paper aims to propose a workflow for automated machine learning-based X-ray image processing, which includes preprocessing phases such as image transformation, feature extraction, selection of training data sets and testing; classification of images into four classes; and efficiency evaluation.

The paper is structured as follows: Methods for data analytics, including types of data mining regularities and data analysis tools, are discussed in Section 2. Section 3 presents knowledge discovery based on biomedical data analytics and a conceptual model for biomedical data analytics. The proposed algorithm for medical image classification selected experimental datasets, and workflow for Covid-19 image analysis are explained in Section 4. Some discussion of the results of this study is concluded in Section 5.

2 Methods for Data Analytics

Data mining is the analysis of stored data concerning extracting knowledge by uncovering hidden relationships between ostensibly unknown and unrelated quantities, [11], [12]. An important feature of data mining is that it provides an opportunity to process multidimensional arrays and extract multidimensional dependencies while automatically revealing exceptional situations - data and dependencies that are not obvious in general patterns. As a result, hypotheses are formed to reveal relations between components and parameters.

The need for a useful yet comprehensive sample of processed data is summarized as follows:

- Unlimited volume of data;
- Great data heterogeneity and variety;
- Specific and understandable results;
- Tools for data processing enable the possibility of simply systematizing and using data.

Modern data mining technologies are built around the idea of patterns or models that reflect the complex relationships between data. These templates are a collection of regularities that execute the selection of data based on specified qualities and present it in ways that are user-friendly and acceptable.

The unpredictable nature of patterns discovered is a key benefit of data mining analysis. As a result, the patterns that are found should show unexpected dependencies in the data that contribute to the so-called hidden meanings. This led to the hypothesis that "raw" data includes far deeper levels of concealed knowledge that can only be exposed after a thorough, in-depth analysis. Table I shows this information. Data mining searches from top to bottom to uncover deeply concealed data that OLAP cannot reveal and analyze. Although OLAP systems and data mining analyses are closely related, there are important differences between the two techniques.

Table 1. Data mining technologies

Type of used technology	The knowledge contained in stored data	Analytical tools used
Top-down	Surface	Language of simple questions
	Shallow	Operational Analytical Processing
Bottom to top	Hidden	Data mining - extraction of the data

The multidimensional analytical technology that enables the efficient use of data kept in a data warehouse is known as OLAP (Online Analytical Processing). It typically comprises tools for interactively analyzing data that has been gathered for a particular user's needs after being retrieved from numerous databases. The ability to exhibit data in several portions makes OLAP technologies substantially more complicated than conventional relational databases.

Although data mining also entails technology that allows for the detection of latent patterns and relationships in various samples, it is also employed for data analysis. Additionally, there are so-called Data Marts, which are essentially local Data Warehouses and hold subsets of aggregated data.

2.1 Types of Data Mining Regularities

Five types of regularities allow data mining analysis to be implemented.

1 Association: It is applied in cases where several events are related to each other.

2 Consistency: If there is a chain of events, it is said to be a sequence.

3 Classification: Signs characterizing the groups in which a given object is included are revealed. This is done by analyzing the classifiable objects and formulating certain rules.

4 Clustering: Homogeneous groups of data having related characteristics are extracted from the classification sets.

5 Prediction: The information found in the data warehouse serves as the foundation for contemporary projections using data mining analysis. They serve as the foundation for the creation of templates that reflect the dynamics of the behavior of the goal indicator and allow for the prediction of future system behavior. A data warehouse is described as a collection of subject-specific, integrated databases with time-sensitive data that are organized to help with decision-making. Both standalone and heavily aggregated data are present in this dataset.

2.2 Data Analysis Tools

There is a wide range of tools to support data analysis. This includes both commonly available algorithms for visualization and machine learning, as well as complex software packages operating based on parallel processors. The use of the most suitable tool for performing data mining analysis is determined by the conditions and objectives of the project. When choosing tools or algorithms, flexibility is very important - how far the chosen strategy can achieve the desired result.

There are numerous stages in the creation of data analytics applications:

Step 1: The project's scope is set, identifying the data that have to be gathered. The project has to be designed to accomplish certain objectives.

2nd step: Create databases. The necessary data can be spread over numerous databases. To get rid of inconsistencies, data from several applications need to be combined and aggregated. The techniques used to create samples from databases based on specific properties should not be altered as data analysis advances. This includes checking for integrity and processing existing values. The accuracy of data mining depends on the quality of the information chosen as its basis.

Step 3: Quantify the data elements. Collaboration is subject-domain experimentation that helps answer questions and isolate the data elements that make the most sense for business needs.

Step 4: Sample data analytics algorithms for determining the relationship between data. It is possible to use several different algorithms to obtain the necessary dependencies. Some of them can be used at the beginning of the process and others at the end. Sometimes several parallel algorithms can also be used to obtain data of different accuracy.

Step 5: Study the relationships manifested at the previous stage of project implementation. At this stage, the help of an expert in the relevant subject area may be needed. He determines whether these ratios are specific or general and indicates in which area the analysis should continue.

Step 6: Presentation of the results in the form of a report revealing the recalculations for all interpreted relations. Such a report is beneficial when the expert can apply a creative approach to analyzing the data and its benefits.

During the development of a data mining project, other factors also influence the type of end application; availability and status of data repositories; the volume of data, its variety, and characteristics.

5. System for visualization of the data obtained from the analysis.

The system occupies an important place. It provides a graphical presentation of the obtained data - graphs, diagrams, schemes, tables, etc. This is done as the visualization system supports an interface, allowing easy association of the analyzed indicators with the various parameters of the diagrams.

3 Biomedical Data Analysis

The process of knowledge discovery based on data analysis consists of the following main stages, presented in Fig. 2, [13], [14], [15]:

Fig. 2: Knowledge Discovery and results interpretation

As in many cases the data are imperfect, containing inconsistencies and abbreviations, they cannot be directly applicable to start the data retrieval process. The rapid increase in the generation rate must also be considered when analyzing data and their size in various academic and scientific applications.

Most of the data collected requires more complex analysis mechanisms. Data preprocessing aims to adapt the data to the requirements set by each data retrieval algorithm, which allows data to be processed that would otherwise be inappropriate.

The data analytics conceptual model is presented in Fig. 3. The overall process related to the extraction and interpretation of data patterns includes the repetition of several steps described below, [16], [17], [18], [19]:

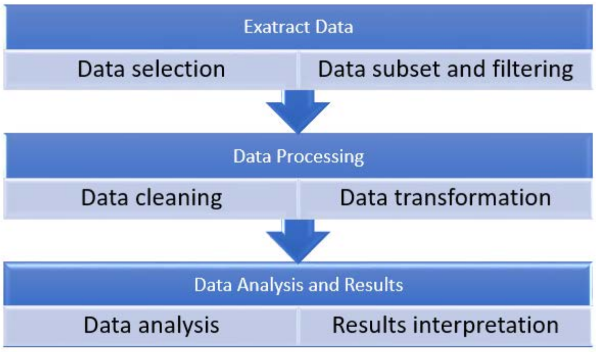


Fig. 3: A conceptual model for biomedical data analytics

- Defining the goal of the process for knowledge discovery - defining the task and corresponding prior knowledge and its application.
- Defining scope, appropriate end-user knowledge, and goals.
- Creating a target dataset: choosing a dataset or selecting variables, sets, or instances of data.

- Filtering and preparation: removal of redundant or negative values; collecting the needed data for modeling; the processing of missing data fields
- Data set simplification by deleting unwanted variables: finding suitable data presentation futures about the task purpose; applying measurement or conversion tools to reduce the number of variables considered.
- Combining the process of data discovery and techniques for data extraction to determine the purpose of the process: classification, clustering, regression, etc.
- Selecting the data extraction algorithm. This includes appropriate models and parameters for the overall process: selecting the method for searching a model in the data; determining appropriate models and parameters; and ensuring compliance of a method for data extraction with the general criteria.
- Data extraction - searching for interesting models such as clustering, regression, classification rules, or trees.
- Interpretation of knowledge derived from models.
- Using knowledge and integrating it into another system for further action.

4 Materials and Methods

4.1 Algorithm for Medical Images Classification

As in many cases the data are imperfect, containing inconsistencies and abbreviations, they cannot be directly applicable to start the data retrieval process. The rapid increase in the rate of generation also has to be taken into account when analyzing data and their size in various academic and scientific applications. Most of the data collected requires more complex analysis mechanisms. Data preprocessing aims to adapt the data to the requirements set by each data retrieval algorithm, which allows data to be processed that would otherwise be inappropriate.

An image analysis algorithm using several machine learning-based algorithms is proposed (Fig. 4). The preprocessing phase is concerned with image transformation, feature extraction, and the selection of training and testing datasets.

The most effective way to use the data is to create a suitable model and address the specific issue once it is determined. As a result of their foundation in mathematical functions and intricate calculations, the majority of machine learning algorithms only accept numbers as input and output. The input representations are consequently transformed into numerical values.

The variables that will be used in the model are chosen once the images being used have been converted. The foundation of analysis and classification models is selecting the appropriate features.

Data preparation, cleansing, and selection; knowledge discovery; decision making and involving findings; and interpreting accurate decisions from observed results are some of the research approaches that are used to extract valuable knowledge from a collection of data. The preprocessing of the training and testing datasets is one of the duties of the preprocessing phase. Preprocessing of data includes feature extraction, feature selection for relevance, and data cleaning for correctness. Because it contains data that can be used to train the system to recognize particular patterns, feature selection is crucial. Using a training dataset and machine learning techniques for classification, the training phase tries to build a repository of models.

The second phase involves analyzing every feature to find and categorize any potential patterns. In the final step, an machine learning algorithm is used to select the best model to capture the behavior or pattern of the data. The machine learning phase utilizes many classification techniques and runs on training and validation datasets. Following the process of feature extraction and data set reduction, the analytical model is designed. Different categorization models are produced as a result, and these models are then utilized to construct an analytical workflow. A validation dataset is used to validate the outcome.

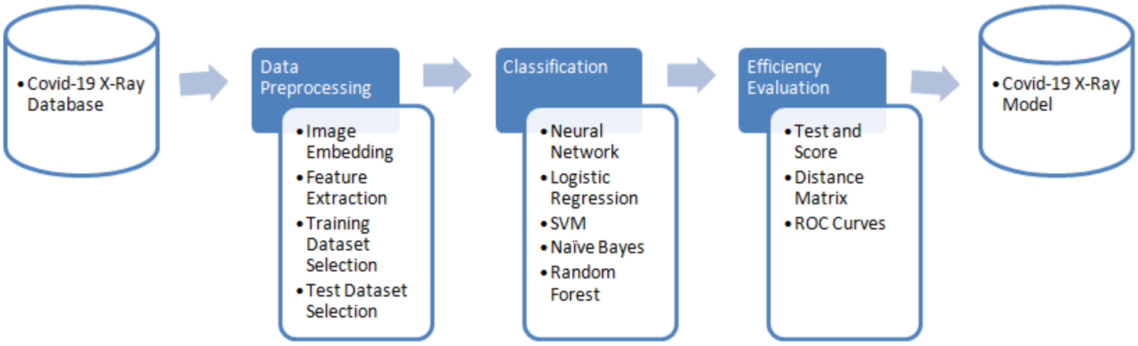


Fig. 4: Algorithm for processing medical images based on machine learning

4.2 Selecting a Dataset

The publicly accessible dataset that was utilized to train and test the suggested approach consists of numerous datasets with four distinct classes, including Normal, COVID-19, Pneumonia, and Lung Opacity, [20], [21], [22], [23]. Different data sets are combined to produce each class. There are 3,616 images in the COVID-19 class total, which were gathered from four various sources. A total of 2473 images were taken from the Padchest dataset, [24]. It represents one of the biggest freely accessible independent databases. Other data sets include 183 chest X-ray images from the German Medical School, [25]. SIRM, GitHub, Kaggle, and Twitter were used to gather 560 chest X-ray images, [26], [27], [28], [29]. Additionally, a dataset, [30], with 400 gathered chest X-ray images is accessible on GitHub.

There are 10192 images of healthy lungs in the dataset of normal lung images, of which 8851 are from RSNA, [31], and 1341 are from Kaggle, [32]. 1345 images had viral pneumonia, [32], while 6012 images were classified as non-COVID-19 lung infection (lung opacity), [31].

4.3 Workflow for Covid-19 Images Analysis

The proposed framework is used to classify the selected medical images by applying different machine-learning algorithms. Five machine learning algorithms are selected to validate the proposed approach: Random Forest, Logistic Regression, Naive Bayes, Neural Network, and SVM. The experimental workflow is implemented using the Orange Data Mining tool.

The Category characteristic is defined as a target for the classification. 66% of the data is designated as the training data set, while the remaining data is the test data set.

The designed workflow for medical image analysis consists of five connected components and is presented in Fig. 5. Four stages of image classification are included: 1) Covid-19 image loading using the component Import Images and image visualization by the Image Viewer component. 2) Feature extraction is performed through the component Inception v.3 Image Embedding. Metadata consists of category, image name, size, width, and height. Additionally, for each image, there are also 2048 features. 3) Distance calculation for similarity assessment using Cosine distance. 4) Hierarchical clustering by categories.

Fig. 6 presents hierarchical clustering. 5) Performance evaluation of accuracy.

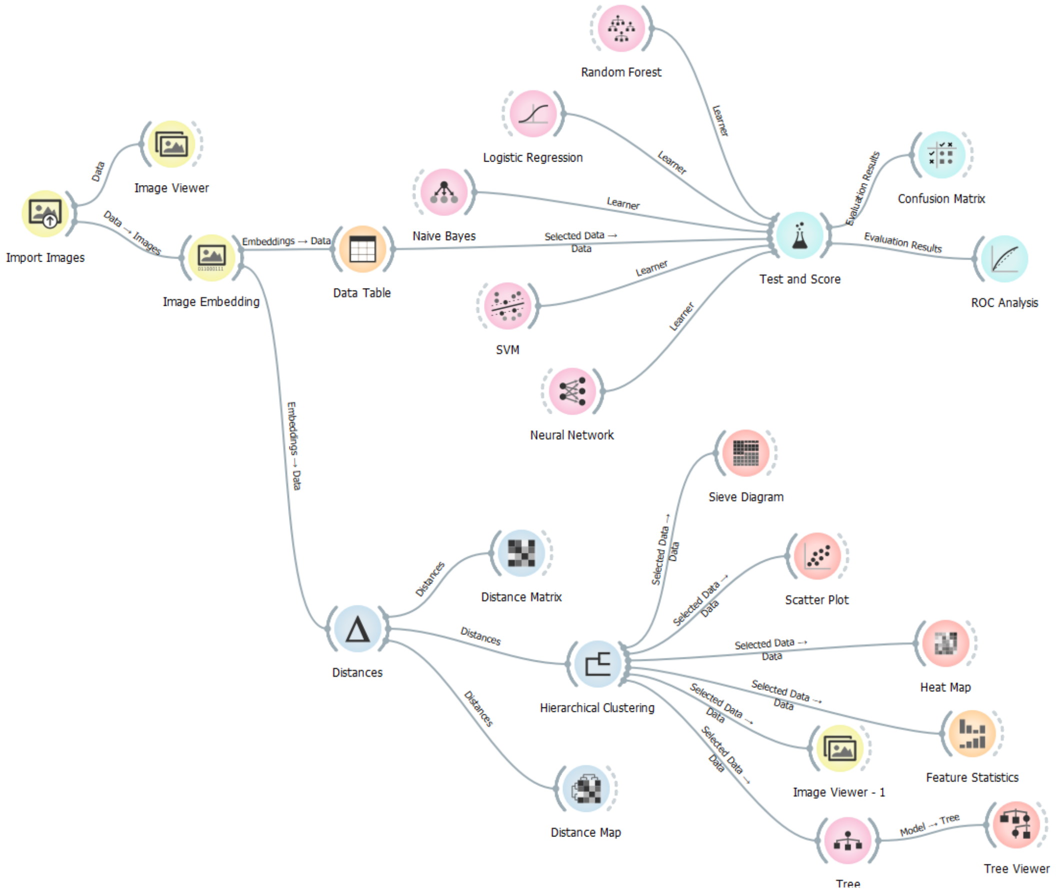


Fig. 5: Workflow for Covid-19 images analysis

The experimental results obtained by the five classification algorithms are shown in Fig. 7.

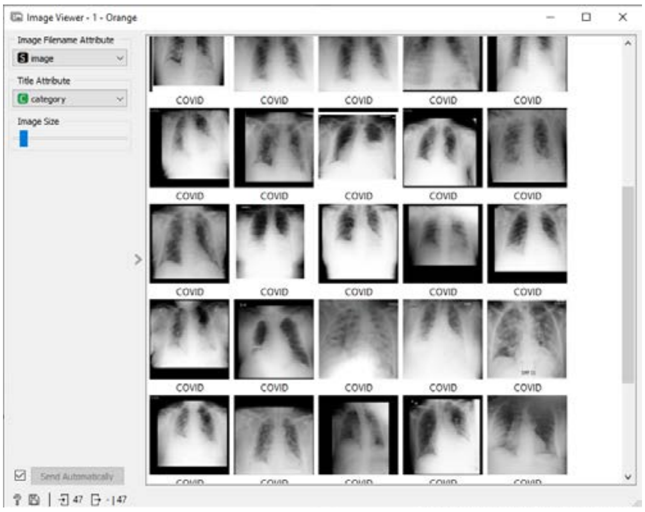
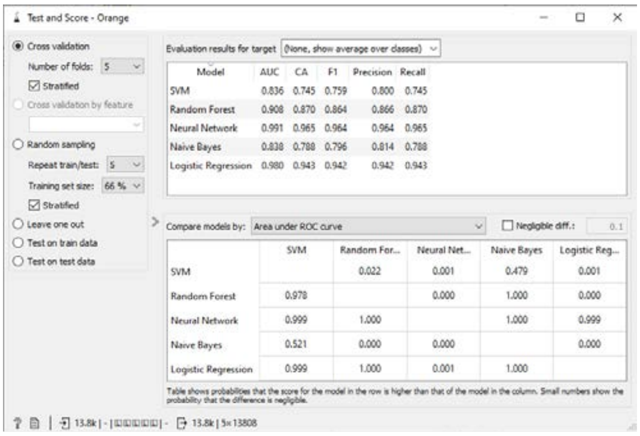


Fig. 6: Graphical representation of hierarchical clustering



● Cross validation
Number of folds: 5
☑ Stratified
○ Cross validation by feature
○ Random sampling
Repeat train/test: 5
Training set size: 66 %
☑ Stratified
○ Leave one out
○ Test on train data
○ Test on test data

Evaluation results for target (None, show average over classes)

Model	AUC	CA	F1	Precision	Recall
SVM	0.836	0.745	0.759	0.800	0.745
Random Forest	0.908	0.870	0.864	0.866	0.870
Neural Network	0.991	0.965	0.964	0.964	0.965
Naive Bayes	0.838	0.788	0.796	0.814	0.788
Logistic Regression	0.980	0.943	0.942	0.942	0.943

Compare models by: Area under ROC curve ☐ Negligible diff.: 0.1

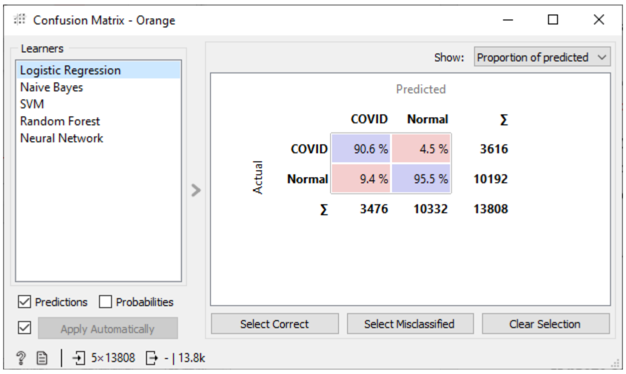
	SVM	Random For...	Neural Net...	Naive Bayes	Logistic Reg...
SVM		0.022	0.001	0.479	0.001
Random Forest	0.978		0.000	1.000	0.000
Neural Network	0.999	1.000		1.000	0.999
Naive Bayes	0.521	0.000	0.000		0.000
Logistic Regression	0.999	1.000	0.001	1.000	

Table shows probabilities that the score for the model in the row is higher than that of the model in the column. Small numbers show the probability that the difference is negligible.

13.8k | - | 13.8k | 5x13808

Fig. 7: Test and score of classification models

The confusion matrix represents the number of cases in the actual and predicted classes. Through the matrix, it is observed that specific cases are not classified correctly. Fig. 8 (a), (b), (c), (d), (e) show the confusion matrix generated by Random Forest, Logistic Regression, Naive Bayes, Neural Network, and SVM models. Accuracy ranges from 0 (lowest accuracy) to 1 (highest accuracy). The Neural Network classification algorithm achieves the best result: 0.964.

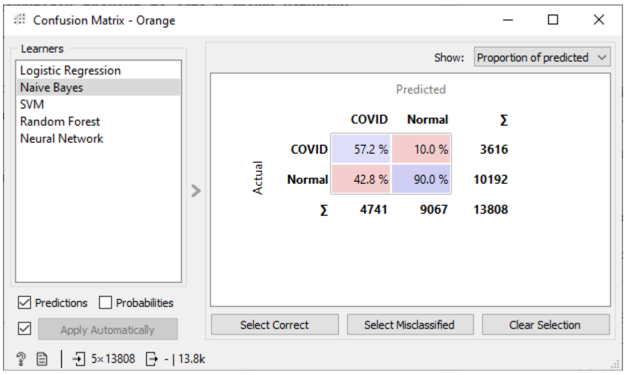


Show: Proportion of predicted

		Predicted		
		COVID	Normal	Σ
Actual	COVID	90.6 %	4.5 %	3616
	Normal	9.4 %	95.5 %	10192
	Σ	3476	10332	13808

☑ Predictions ☐ Probabilities
☑ Apply Automatically

(a)

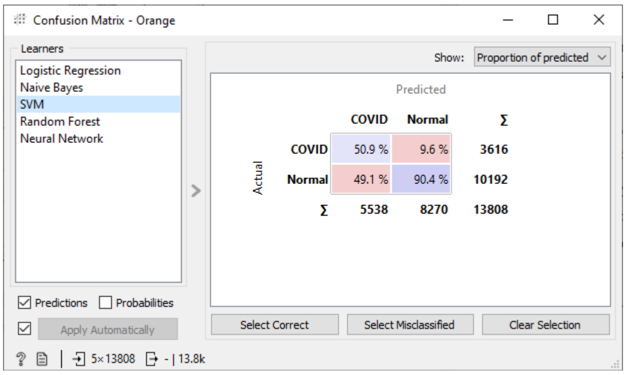


Show: Proportion of predicted

		Predicted		
		COVID	Normal	Σ
Actual	COVID	57.2 %	10.0 %	3616
	Normal	42.8 %	90.0 %	10192
	Σ	4741	9067	13808

☑ Predictions ☐ Probabilities
☑ Apply Automatically

(b)

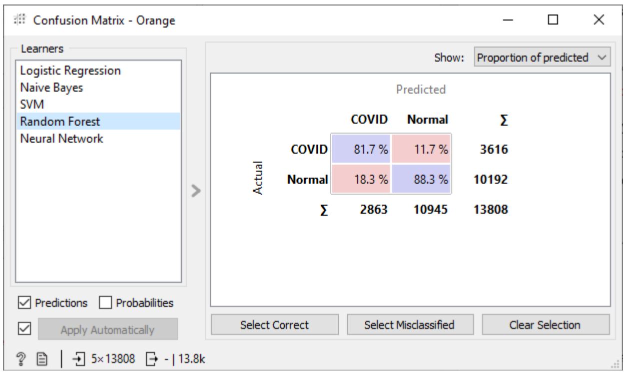


Show: Proportion of predicted

		Predicted		
		COVID	Normal	Σ
Actual	COVID	50.9 %	9.6 %	3616
	Normal	49.1 %	90.4 %	10192
	Σ	5538	8270	13808

☑ Predictions ☐ Probabilities
☑ Apply Automatically

(c)

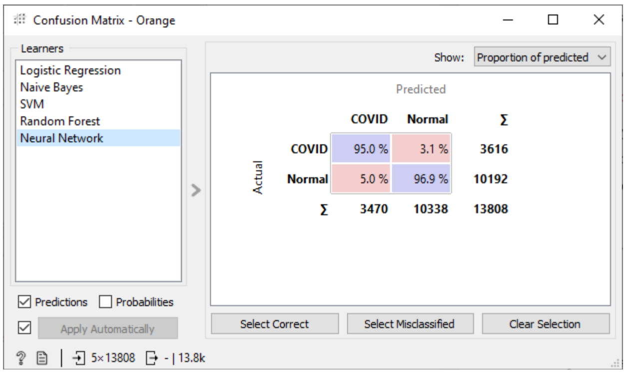


Show: Proportion of predicted

		Predicted		
		COVID	Normal	Σ
Actual	COVID	81.7 %	11.7 %	3616
	Normal	18.3 %	88.3 %	10192
	Σ	2863	10945	13808

☑ Predictions ☐ Probabilities
☑ Apply Automatically

(d)



Show: Proportion of predicted

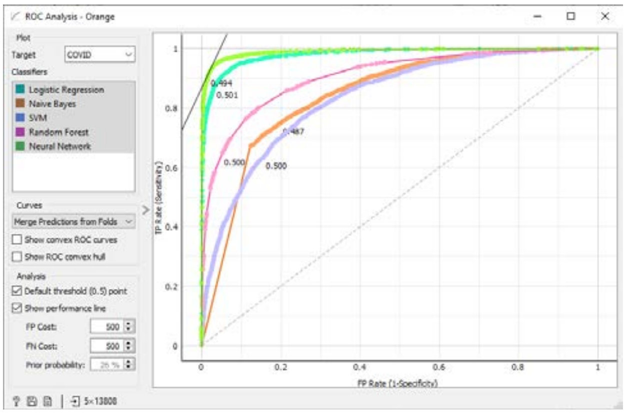
		Predicted		
		COVID	Normal	Σ
Actual	COVID	95.0 %	3.1 %	3616
	Normal	5.0 %	96.9 %	10192
	Σ	3470	10338	13808

☑ Predictions ☐ Probabilities
☑ Apply Automatically

e)

Fig. 8: Generated classification matrix by methods: (a) Logistic Regression; (b), Naïve Bayes; (c) SVM; d) Random Forest; e) Neural Network

The ROC analysis compares the true positive rate and the false positive rate. The results of testing classification algorithms are input for ROC analysis. The widget displays the convex hull and ROC curves for each of the tested models. It acts as a benchmark for contrasting categorization models. On the graph, the true positive rate is shown on the y-axis (sensitivity), while the false positive rate is shown on the x-axis (1-specificity; a chance that target=1 when true value=0). The classifier is more accurate the more closely the curve resembles the left-hand border and then the top border of the ROC space. The widget can also choose the ideal classifier and threshold based on the costs of false positives and false negatives. Fig. 9 presents the ROC curves through which the classifiers can be observed and the results for the selected classification models can be compared.



Target: COVID
Merge Predictions from Folds
☐ Show convex ROC curves
☐ Show ROC convex hull
☑ Default threshold (0.5) point
☑ Show performance line
FP Cost: 500
FN Cost: 500
Prior probability: 25 %

Fig. 9: ROC analysis of classification models

5 Conclusion

This paper presents an approach based on machine learning algorithms for the classification of X-ray images. An algorithm for X-ray medical image processing based on machine learning is proposed. The preprocessing phase is concerned with image transformation, feature extraction, and the selection of training and testing datasets.

Regarding system testing and validation purposes, a workflow is developed to classify X-ray images and determine accuracy and probability in medical data analysis.

An open-source X-ray image dataset was utilized to train and test the suggested approach consisting of numerous datasets with four distinct classes, including Normal, COVID-19, Pneumonia, and Lung Opacity.

Experiments are performed based on Naive Bayes, Random Forest, Logistic Regression, Neural Network, and SVM methods and are aimed at accuracy and probability in the analysis of medical images.

The analysis done shows the best results in the case of the Neural Networks classification algorithm and can be assumed to be the most reliable in comparison with the results in the cases of Random Forest, Logistic Regression, Naive Bayes, and SVM.

Acknowledgment:
The research presented in this paper is financially supported by the Bulgarian Ministry of Education, National Science Fund, grant KP-06-N37/24.

References:

[1] Mallappallil M., et al, A Review of Big Data and Medical Research, *SAGE Open Med*, vol 8, 2020, doi: 10.1177/2050312120934839.

[2] Belle A., et al, Big Data Analytics in Healthcare. *Biomed Res Int*, 2015, doi: 10.1155/2015/370194.

[3] Esfandiari N., et al, Knowledge Discovery in Medicine: Current Issue and Future Trend, *Expert Systems with Applications,* vol. 41, 2014, https://doi.org/10.1016/j.eswa.2014.01.011.

[4] Soffer S., et al, Convolutional Neural Networks for Radiologic Images: A Radiologist's Guide, *Radiology*, 2019, Vol. 290, No. 3, https://doi.org/10.1148/radiol.2018180547.

[5] Borkowski A., et al, Using Artificial Intelligence for COVID-19 Chest X-ray Diagnosis, Federal practitioner: for the health care professionals of the VA, DoD, and PHS, 2020, https://doi.org/10.12788/fp.0045.

[6] Reshi A., et al, An Efficient CNN Model for COVID-19 Disease Detection Based on X-Ray Image Classification, *Complexity*, vol. 2021, https://doi.org/10.1155/2021/6621607.

[7] Wang S., et al, A deep learning algorithm using CT Images to Screen for Corona Virus Disease (COVID-19). *Eur Radiol* 31, 6096–6104, 2021, https://doi.org/10.1007/s00330-021-07715-1.

[8] Arias-Garzón D., et al, COVID-19 Detection in X-ray Images Using Convolutional Neural Networks, *Machine Learning with Applications*, Volume 6, 2021, 100138, ISSN 2666-8270, https://doi.org/10.1016/j.mlwa.2021.100138.

[9] Ozturk T., et al, Automated Detection of COVID-19 Cases Using Deep Neural Networks with X-ray Images, *Comput. Biol. Med.*, 2020, doi: 10.1016/j.compbiomed.2020.103792.

[10] Imad M., et al, COVID-19 Classification Based on Chest X-Ray Images Using Machine Learning Techniques, *J. Comput. Sci. Technol. Stud.*, vol. 2, no. 2, pp. 01–11, 2020.

[11] Hlaing K. S., Thaw Y. M. K. K., Applications, Techniques and Trends of Data Mining and Knowledge Discovery Database, *International Journal of Trend in Scientific Research and Development*, vol. 3, 2019, pp. 1604-1606.

[12] Pushp, Chand S., Knowledge Discovery and Data Mining for Intelligent Business Solutions, *Advances in Data and Information Sciences. Lecture Notes in Networks and Systems*, vol. 318, 2022, https://doi.org/10.1007/978-981-16-5689-7_18.

[13] Pareek M., Bhari P., A Review Report on Knowledge Discovery in Databases and Various Techniques of Data Mining, *Open Access International Journal of Science and Engineering*, 2020. pp. 79-82.

[14] Parihar A., Sharma S., Knowledge Discovery and Data Mining Healthcare, *International Journal of Information Technology Insights & Transformations*, vol. 4, Issue 1, 2020.

[15] Borovska P., Gancheva V., Georgiev I., Platform for Adaptive Knowledge Discovery and Decision Making Based on Big Genomics Data Analytics, *Bioinformatics and Biomedical Engineering, Lecture Notes in Computer Science*, vol. 11466, Springer, pp. 297–308, https://doi.org/10.1007/978-3-030-17935-9_27.

[16] Janiesch C., Zschech P., Heinrich K., Machine Learning and Deep Learning, *Electronic Markets*, vol. 31, pp. 685-695.

[17] Sun S., et al, A Survey of Optimization Methods from a Machine Learning Perspective, *IEEE Transactions on Cybernetics*, vol. 50, no. 8, 2019, pp. 3668-3681.

[18] Sarker I. H., Machine Learning: Algorithms, Real-World Applications and Research Directions, *SN Computer Science*, 2021, 1-21, https://doi.org/10.1007/s42979-021-00592-x.

[19] Mahesh B., Machine Learning Algorithms - a Review, *International Journal of Science and Research*, vol. 9, 2020, pp. 381-386.

[20] Chowdhury M., et al, Can AI Help in Screening Viral and COVID-19 Pneumonia?, *IEEE Access*, Vol. 8, 2020, pp. 132665 - 132676.

[21] Khan E., et al, Chest X-Ray Classification for the Detection of COVID-19 Using Deep Learning Techniques, *Sensors*, 2022, doi: 10.3390/s22031211.

[22] Rahman T., et al, Exploring the Effect of Image Enhancement Techniques on COVID-19 Detection using Chest X-Ray Images, *arXiv*, https://doi.org/10.48550/arXiv.2012.02238.

[23] COVID-19 Radiography Database, https://www.kaggle.com/tawsifurrahman/covid19-radiography-database (accessed 17 April 2023).

[24] BIMCV-COVID19 Database, https://bimcv.cipf.es/bimcv-projects/bimcv-covid19/ (accessed 17 April 2023).

[25] COVID-19-Image-Repository, https://github.com/ml-workgroup/covid-19-image-repository/tree/master/png (accessed 17 April 2023).

[26] Chen R., et al, Risk Factors of Fatal Outcome in Hospitalized Subjects with Coronavirus Disease 2019 from a Nationwide Analysis in China, *Chest*, 2020, 158, 97–105, doi: 10.1016/j.chest.2020.04.010.

[27] Weng Z., et al, ANDC: An Early Warning Score to Predict Mortality Risk for Patients with Coronavirus Disease 2019, *Journal of Translational Medicine*, 328, 2020.

[28] Liu L., et al, Neutrophil-to-Lymphocyte Ratio Predicts Severe Illness Patients with 2019 Novel Coronavirus in the Early Stage, *Journal of Translational Medicine*, 2020, doi: 10.1186/s12967-020-02374-0.

[29] Huang I., Pranata I., Lymphopenia in Severe Coronavirus Disease - 2019 (COVID-19): Systematic Review and Meta-Analysis, *Journal Intensive Care*, 2020, https://doi.org/10.1186/s40560-020-00453-4.

[30] COVID-CXNet Dataset, https://github.com/armiro/COVID-CXNet (accessed 17 April 2023).

[31] RSNA Pneumonia Detection Challenge, ttps://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data (accessed 17 April 2023).

[32] Chest X-ray Images (Pneumonia) Dataset, https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia (accessed 17 April 2023).

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

-Veska Gancheva proposed the methodology and algorithm.

-Ivaylo Georgie investigated methods for data analytics.

-Violeta Todorova executed the experiments.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

The research presented in this paper is financially supported by the Bulgarian Ministry of Education, National Science Fund, grant KP-06-N37/24.

Conflict of Interest

The authors have no conflicts of interest to declare.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)