JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 04 Issue 02
2023

Research Article
https://doi.org/10.56979/402/2023

Predicting the Metastasis Ability of Prostate Cancer using Machine Learning Classifiers

Ahmad Naeem[1], Ali Haider Khan[1,2*], Salah u din Ayubi[3], and Hassaan Malik[1]

[1]Department of Computer Science, School of Science & Technology, University of Management and Technology, Lahore 54000, Pakistan.
[2]Department of Software Engineering, Faculty of Computer Science, Lahore Garrison University, Lahore, 54000, Pakistan.
[3]Department of Information Technology, Faculty of Computer Science, Lahore Garrison University, Lahore, 54000, Pakistan.
[*]Corresponding Author: Ali Haider Khan. Email: ali.khan@lgu.edu.pk

Received: December 29, 2022 **Accepted:** March 16, 2023 **Published:** March 29, 2023.

Abstract: Patients with prostate cancer (PCA) are more vulnerable to metastasis, which is the disease's most devastating result and the primary reason for mortality. It is still not possible to accurately anticipate whether or not locally advanced PCA will spread. In this work, potential biomarkers are identified by using Machine learning, which compares the gene expressions of metastatic and local prostate cancer by identifying the differentially expressed genes (DEGs) and the molecular pathways associated with the metastasis development of prostate cancer. Two gene profiles (GSE32269 and GSE 6919) are downloaded from the Gene Expression Omnibus collection, which contains a total of 226 tissue samples (69 metastatic, 81 normal prostates, and 76 localized PCA). A fine-tuned Support vector machine(SVM) for feature selection and classification is used, which is employed to analyze gene activity and select vital biomarkers. Moreover, this study examines the genomic activity and determines the key gene that is essential in distinguishing between localized and metastatic PCA.

Keywords: Prostate Cancer; Metastasis; Machine Learning; SVM; Classification; Biomarkers.

1. Introduction

The most commonly diagnosed cancer in men is prostate cancer and it is the fifth leading cause of death in men [1-5]. In prostate cancer, a malignant tumor is created in the urinary system of males belonging to American and European populations [2]. The semen fluid is produced by the prostate gland which presents below the bladder of a man. Most old men suffer from prostate cancer. It is rare in men which are less than 40 years of age. Males over the age of 65 have a high risk of prostate cancer if they have a family history of prostate cancer. However, prostate cancer is a relatively slow-growing disease and sometimes does not show any symptoms in the early stages. In the advanced stages, prostate cancer symptoms might include urinary problems, pain in the lower back, and ejaculation pain [3]. The state of cancer is determined by the rate at which cancer spread and how distinct it is from the tissues around it. Tumor cell growth, migration, and invasion lead to metastases. In general, localized prostate cancer initially spread to lymph nodes, then it spread to bones, lungs, or liver [4-11]. Bone metastases are identified in around 70% of patients with advanced prostate cancer [10]. Therefore, the last thing which doctors want to see is cancer metastasis.

The five-year local prostate cancer has a survival rate of 100% whereas, prostate cancer with metastatic reduces the survival rate to 30%. Extensive investigations revealed that the biomolecules such as miRNA, enzymes, and in some cases abnormal glucose metabolisms are the cause of the development and metastasis of prostate cancer [9]. The prediction of prostate cancer remains unfinished because each cascade element is not holistically replicated in the metastasis of prostate cancer [8]. Digital rectal test (DRE), prostate-specific antigen (PSA) blood test, and ultrasonography are traditionally used to detect prostate cancer.

All these methods have low sensitivity and specificity which are not up to the mark with medical standards [12-16]. Magnetic resonance imaging (MRI) for prostate cancer is unable to recognize 12% of cancer cases. However, before the symptoms arise these tests are also utilized for the screening of prostate cancer. Moreover, MRI or biopsy may be required if the results are abnormal. Whereas the treatment is typically based on the cancer stage [17].

To increase the treatment efficiency, gene biomarkers are used to find the location of prostate cancer. The unavailability of biomarkers leads to a poor prediction of the metastatic potential of local prostate cancer [18-24]. Moreover, it is important to understand the biological differences between localized and metastatic prostate cancer for the development of new biomarkers which helps in prostate cancer prognosis and treatment. The cause of disease development can easily understand by the analysis of biomarkers [25-32]. Applying artificial intelligence (AI) to help detect the metastatic nature of PCA. Researchers have primarily focused on the gene dataset to enhance the metastasis detection of PCA [33].

Machine learning algorithms have effectively been implemented to identify the gene biomarkers of prostate cancer. The advancement of next-generation sequencing (NGS) technology attracts researchers to identify the genomic biomarker for prostate cancer [34]. NGS increases the accuracy of gene relationship detection and provides a detailed view of cancer cell gene transcription activities [35]. Preprocessing of data is required because NGS produces a lot of data with certain abnormalities. Gene expression data provides a deep insight into the genomic activities of tumor tissues which helps in a better understanding of disease development [34].

In this work, we apply a machine-learning algorithm (ML) to predict the metastasis of prostate cancer. To classify the localized and metastasis PCA, a fine-tuned support vector machine (SVM) is applied to two gene profiles GSE 6919 and GSE 32269.

2. Materials and Methods

This section includes the proposed methodology for our study for the development of a biomarker to identify localized and metastatic PCA.

2.1. Microarray data

The GEO database is used for the download of the GSE 32269 and GSE 6919 gene expression dataset for this study [13]. Each dataset contains the primary prostate cancer samples as well as metastasis prostate cancer. Moreover, sample ID, gene symbol, and entering gene ID are included in the platform file. The raw gene expression data set file type is CEL. GSE32269 dataset contains 55 samples, 16 samples belong to localized prostate cancer whereas 39 samples belong to metastasis prostate cancer. The GPL96 (Affymetrix Human Genome U133A Array) platform is used to build this dataset. The GSE6919 dataset contains 171 samples, 65 samples belong to primary localized prostate cancer, 81 samples belong to normal prostate tissues and 25 samples belong to metastasis prostate cancer. This data set is constructed using the Affymetrix Human Genome U95 Version 2.0 Array (GPL8300) platform [14].

2.2. Data Preprocessing

To develop an equal contribution of features the genes expression are scaled by the technique of data normalization. The Affy package of R language is used to normalize the raw CEL data. The corresponding gene symbol is used to represent each probe ID of the expression matrix in the annotation file. The R language is used to calculate the average value when multiple probes correlate to the same gene. The lemma R package is utilized to filter the genes of each dataset. Differentially expressed genes have been considered for genes when the P-value < 0.05 and $|\log2\text{fold changes (FC)}| > 1$ are observed [14]. The quality of data plays a significant role in the ML-based classifier [16].

2.3. Data Sampling

The class imbalance problem is observed in the dataset, the metastasis occurrence in the adrenal gland, lungs, and kidney, recurrent in the prostate and left inguinal lymph node is one whereas, the retroperitoneal and para-aortic lymph node is three. The metastasis occurrence of the liver is five and the paratracheal lymph node is eight. The metastasis occurrence of bone is thirty-nine. To identify the most effective solution for our datasets, multiple resampling methods have been implemented and tested [15]. To ensure that each group has an equivalent number of samples, oversampling simply adds the duplicates to the minority groups. Oversampling, on the other hand, helps even out the classes and improves the classifier's accuracy, but it also has the drawback of being prone to overfitting. Oversampling of underrepresented groups is accomplished with the Synthetic Minority Oversampling Method (SMOTE) while undersampling of dominant groups is achieved with the Neighborhood Cleaning Rule (NCL). SMOTE and NCL worked better

than any other strategy [16-17] for dealing with unbalanced data. If a group's sample doesn't share characteristics with at least two of its three nearest neighbors, NCL throws out the group's sample. By employing the feature vector that connects both samples, SMOTE generates a third synthetic sample to go along with the two original samples (see Figure 1). A new sample's precise position is determined by measuring the distance between two existing samples and then multiplying that result by a random number between 0 and 1 [17].

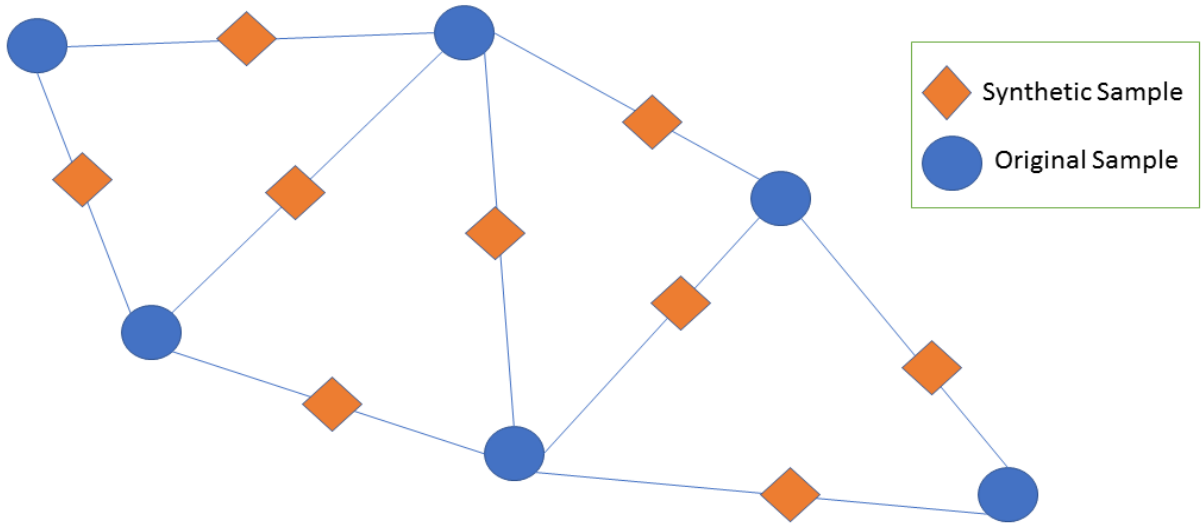

Figure 1. Synthetic minority oversampling technique (SMOTE)

2.4. Feature Selection

The curse of dimensionality becomes an issue when dealing with a large variety of features. Machine learning techniques are used to speed up the classification process and cut down on the amount of identified features. All genes are divided into groups according to the largest information gain (IG) against each category using the information gain (IG) approach for selecting features. Attributes with the lowest scores are thrown out, while those with the highest scores are chosen. In this research, the IG considers all of the characteristics [19] The IG of feature J concerning class K may be found by using the following formula:

$$IG(K,J) = H(K) - H(K|J) \tag{1}$$

where,

$$H(K) = -\sum_{k \in K} p(k) log2(p(k)) \tag{2}$$

$$H(K|J) = -\sum_{j \in J} p(j) \sum_{k \in J} p(b|j) log2\big(p(K|j)\big) \tag{3}$$

Here, *H(K)* shows the entropy of class K and *H(K|J)* shows the qualified entropy of K given by J. This stage is used to select the best set of attributes (genes) to ensure the right classification between the various classes.

2.5. Machine learning Models

Appropriate feature selection plays a significant role to measure classification accuracy. A fine-tuned SVM is utilized to evaluate the performance of a biomarker that identifies metastatic and localized PCA.

2.5.1 Support Vector Machine (SVM)

The SVM is an ML classifier that is mostly utilized for regression and classification problems. It is a linear model which figures out hyperplane for n-dimensional data. SVM split the data within the decision limit.

SVM is a linear model which is mostly used for classification and regression problems [19]. SVM is one of the most widely used ML algorithms to figure out the best hyperplane for the n-dimensional (D)

space of data [26]. The work of SVM is to split the data points according to space within a decision boundary. The following equations (4) and (5) represent the positive (+) and negative (-) relation used for dividing the space between the decision boundaries.

$$x * y + a > 0 \quad (4)$$

$$x * y + a < 0 \quad (5)$$

Whereas x represents the value of the vector perpendicular to the median, the value of the unknown vector is represented by the y and the constant is represented by the a.

A novel framework is proposed in this study for the detection of localized and metastatic PCA. Both the gene expression datasets are merged, and the proposed model is applied to create a biomarker for the classification of local and metastatic PCA. In the initial step data, pre-processing is applied to the gene expression datasets. Data is normalized at this stage. The data imbalance issue is resolved by using the SMOTE and NCL techniques. Afterward, the data is split into three parts, 70% of the data is used for training, 20% is used for testing and 10 % is used for validation purposes. To train the proposed SVM random gene instances are selected with k-fold cross-validation. The SVM passes the data input as an argument kernel by utilizing a support vector classifier (SVC) as a linear kernel. To classify the data, SVM uses hyperplanes. The hinge cost function is used to evaluate the margin between hyperplanes and data points. To prevent the model overfitting cross-validation is performed, which helps to observe the learning status of the model at the time of training. The SVM is fine-tuned by using the stochastic gradient descent (SGD) on the initial learning rate (0.0006, 0.06 & 0.2) with a momentum value of 0.8. a total of 4 epochs are used to execute the SVM. The accuracy of the model is observed after every 0.5 epochs, if the model seems to be overfitting and its accuracy is not increasing then the learning rate is reduced by a factor of 0.2 [24]. Figure 2, represents the proposed method.

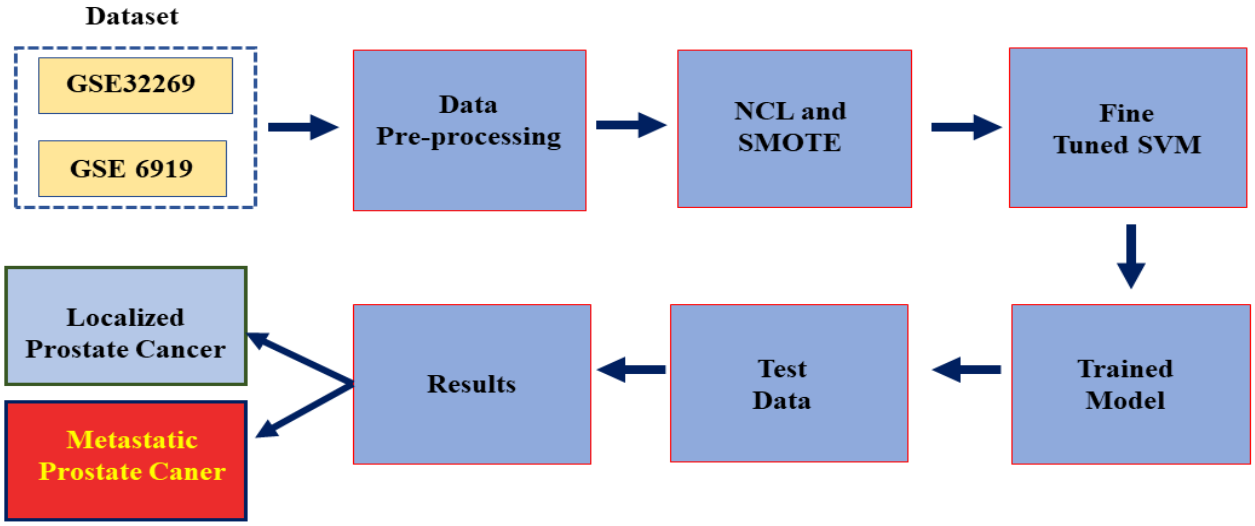

Figure 2. Proposed Methodology

2.6. Evaluation Criteria

The test data were used to evaluate the ML models. The performance of the biomarker for the classification of localized and metastatic PCA is calculated using the confusion matrix. The correct instances in the confusion matrix are represented by the true positive (TP) and false positive (FP), whereas the incorrect instances are represented by the true negative (TN) and false negative (FN). Based on these instances, the effectiveness of the machine learning algorithm is measured in terms of accuracy, sensitivity, specificity, and f1 score [16].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (7)$$

$$Specificity = \frac{TN}{TN + FP} \quad (8)$$

$$F1\ score = 2 * (\frac{Sen * Spec}{Sen + Spec}) \quad (9)$$

3. Results and Discussion

The training, testing, and validation phenomenon measure the prediction performance of the proposed method on two datasets (GSE 6919 and GSE 32269). For the experimentation, SVM is implemented in Python language using SkLearn. The five epochs are used in the proposed SVM. The proposed model achieves a training and validation accuracy of 0.97 and 0.96 respectively. Moreover, the training loss of 0.03 and 0.04 of validation loss is proposed by the proposed method. These findings present that the proposed SVM is trained well on the given dataset. The hyperparameters for the SVM are the sigmoid function, linear kernel, and regularization parameters. After extracting features from the dataset, the SVM model was trained. The grid search techniques are used to fine-tune the value of the hyperparameter of the proposed model.

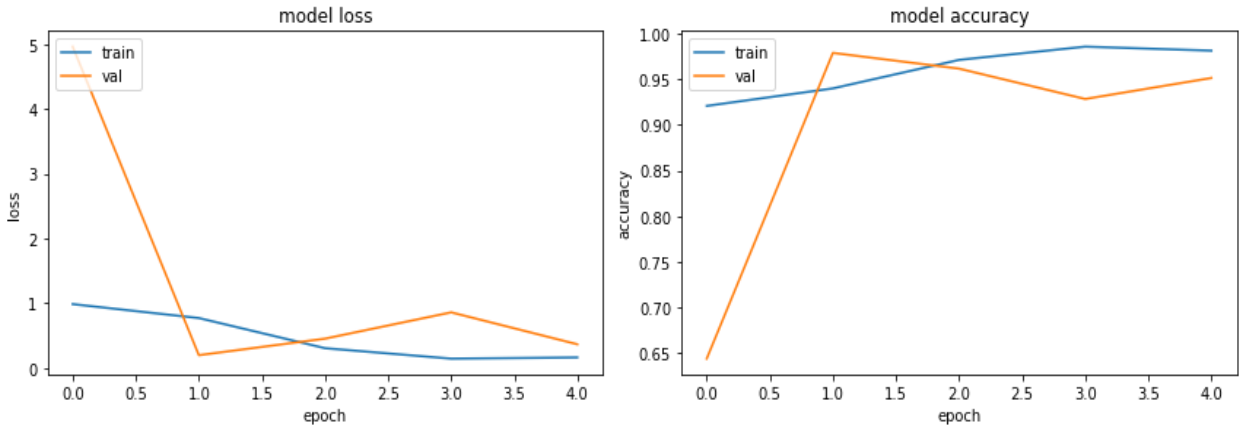

Figure 3. Proposed SVM Training-validation accuracy and loss.

The one versus all approach is used to solve the classification problem. The classification method is evaluated by using Accuracy, specificity, and sensitivity. Support vector machine algorithm is applied for the classification of localized and metastasis PCA based on feature selection. The proposed method achieves the highest accuracy and precision. The 10-fold cross-validation is used in our model.

4. Conclusions

In this study, a machine learning-based model identifies the gene activity that helps in the detection and understanding of the metastatic nature of prostate cancer with high accuracy. Moreover, these genes could be potential indicators of metastatic PCA and therapeutic targets. We have uncovered genes that can discriminate localized and metastatic prostate cancer with great accuracy. Moreover, other types of cancer and clinical problems can be investigated by employing the proposed method. This study provides potential biomarkers that work as an alternative for painful biopsies and misleading image scans. However, more investigations are still necessary to validate these results.

Data Availability Statement: Publicly archived datasets used during the study.
Conflicts of Interest: "The authors declare no conflict of interest."

References

1. Guo, Z., Zhang, R., Li, Q., Liu, X., Nemoto, D., Togashi, K., ... & Zhu, X. (2020, April). Reduce false-positive rate by active learning for automatic polyp detection in colonoscopy videos. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (pp. 1655-1658). IEEE.
2. Siegel, D. A., O'Neil, M. E., Richards, T. B., Dowling, N. F., & Weir, H. K. (2020). Prostate cancer incidence and survival, by stage and race/ethnicity—United States, 2001–2017. *Morbidity and Mortality Weekly Report*, *69*(41), 1473.
3. Moschini, M., Zaffuto, E., Karakiewicz, P., Mattei, A., Gandaglia, G., Fossati, N., ... & Shariat, S. F. (2019). The effect of androgen deprivation treatment on subsequent risk of bladder cancer diagnosis in male patients treated for prostate cancer. *World Journal of Urology*, *37*, 1127-1135.
4. Zhou, C. K., Pfeiffer, R. M., Cleary, S. D., Hoffman, H. J., Levine, P. H., Chu, L. W., ... & Cook, M. B. (2015). Relationship between male pattern baldness and the risk of aggressive prostate cancer: an analysis of the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. *Journal of clinical oncology*, *33*(5), 419.
5. Drost, F. J. H., Osses, D. F., Nieboer, D., Steyerberg, E. W., Bangma, C. H., Roobol, M. J., & Schoots, I. G. (2019). Prostate MRI, with or without MRI-targeted biopsy, and systematic biopsy for detecting prostate cancer. *Cochrane Database of Systematic Reviews*, (4).
6. Carroll, P. R., Parsons, J. K., Andriole, G., Bahnson, R. R., Castle, E. P., Catalona, W. J., ... & Freedman-Cass, D. A. (2016). NCCN guidelines insights: prostate cancer early detection, version 2.2016. *Journal of the National Comprehensive Cancer Network*, *14*(5), 509-519.
7. Wang, G., Zhao, D., Spring, D. J., & DePinho, R. A. (2018). Genetics and biology of prostate cancer. *Genes & development*, *32*(17-18), 1105-1140.
8. Su, S. C., Hsieh, M. J., Yang, W. E., Chung, W. H., Reiter, R. J., & Yang, S. F. (2017). Cancer metastasis: Mechanisms of inhibition by melatonin. *Journal of pineal research*, *62*(1), e12370.
9. Pan, D., Jia, Z., Li, W., & Dou, Z. (2019). The targeting of MTDH by miR-145-5p or miR-145-3p is associated with prognosis and regulates the growth and metastasis of prostate cancer cells. *International journal of oncology*, *54*(6), 1955-1968.
10. Berish, R. B., Ali, A. N., Telmer, P. G., Ronald, J. A., & Leong, H. S. (2018). Translational models of prostate cancer bone metastasis. *Nature Reviews Urology*, *15*(7), 403-421.
11. Alkhateeb, A., Rezaeian, I., Singireddy, S., Cavallo-Medved, D., Porter, L. A., & Rueda, L. (2019). Transcriptomics signature from next-generation sequencing data reveals new transcriptomic biomarkers related to prostate cancer. *Cancer informatics*, *18*, 1176935119835522.
12. Hamzeh, O., Alkhateeb, A., Zheng, J., Kandalam, S., & Rueda, L. (2020). Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data. *BMC bioinformatics*, *21*(2), 1-10.
13. Naji, L., Randhawa, H., Sohani, Z., Dennis, B., Lautenbach, D., Kavanagh, O., ... & Profetto, J. (2018). Digital rectal examination for prostate cancer screening in primary care: a systematic review and meta-analysis. *The Annals of Family Medicine*, *16*(2), 149-154.
14. Gautier, L., Cope, L., Bolstad, B. M., & Irizarry, R. A. (2004). affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, *20*(3), 307-315.
15. Junsomboon, N., & Phienthrakul, T. (2017, February). Combining over-sampling and under-sampling techniques for imbalance dataset. In *Proceedings of the 9th International Conference on Machine Learning and Computing* (pp. 243-247).
16. Bae, S. Y., Lee, J., Jeong, J., Lim, C., & Choi, J. (2021). Effective data-balancing methods for class-imbalanced genotoxicity datasets using machine learning algorithms and molecular fingerprints. *Computational Toxicology*, *20*, 100178.
17. Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020, April). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)* (pp. 243-248). IEEE.
18. Guo, L., Lin, M., Cheng, Z., Chen, Y., Huang, Y., & Xu, K. (2019). Identification of key genes and multiple molecular pathways of metastatic process in prostate cancer. *PeerJ*, *7*, e7899.
19. Kang, C., Huo, Y., Xin, L., Tian, B., & Yu, B. (2019). Feature selection and tumor classification for microarray data using relaxed Lasso and generalized multi-class support vector machine. *Journal of theoretical biology*, *463*, 77-91.
20. Naeem, A., Farooq, M. S., Khelifi, A., & Abid, A. (2020). Malignant melanoma classification using deep learning: datasets, performance measurements, challenges and opportunities. *IEEE Access*, *8*, 110575-110597.
21. Naeem, A., Anees, T., Naqvi, R. A., & Loh, W. K. (2022). A comprehensive analysis of recent deep and federated-learning-based methodologies for brain tumor diagnosis. *Journal of Personalized Medicine*, *12*(2), 275.
22. Naeem, A., Anees, T., Fiza, M., Naqvi, R. A., & Lee, S. W. (2022). SCDNet: A Deep Learning-Based Framework for the Multiclassification of Skin Cancer Using Dermoscopy Images. *Sensors*, *22*(15), 5652.
23. Malik, H., Anees, T., Din, M., & Naeem, A. (2022). CDC_Net: multi-classification convolutional neural network model for detection of COVID-19, pneumothorax, pneumonia, lung Cancer, and tuberculosis using chest X-rays. *Multimedia Tools and Applications*, 1-26.
24. Malik, H., Naeem, A., Naqvi, R. A., & Loh, W. K. (2023). DMFL_Net: A Federated Learning-Based Framework for the Classification of COVID-19 from Multiple Chest Diseases Using X-rays. *Sensors*, *23*(2), 743.

25. Naeem, A., Anees, T., Ahmed, K. T., Naqvi, R. A., Ahmad, S., & Whangbo, T. (2022). Deep learned vectors' formation using auto-correlation, scaling, and derivations with CNN for complex and huge image retrieval. *Complex & Intelligent Systems*, 1-23.
26. Malik, H., Anees, T., Naeem, A., Naqvi, R. A., & Loh, W. K. (2023). Blockchain-Federated and Deep-Learning-Based Ensembling of Capsule Network with Incremental Extreme Learning Machines for Classification of COVID-19 Using CT Scans. *Bioengineering*, *10*(2), 203.
27. Ahmed, K. T., Tahir, I., & Naeem, A. (2017). Features from Accelerated Segment Test with BoW for Effective Image Retrieval. *IJCSIS*, *15*(4).
28. Malik, H., Farooq, M. S., Khelifi, A., Abid, A., Qureshi, J. N., & Hussain, M. (2020). A comparison of transfer learning performance versus health experts in disease diagnosis from medical imaging. *IEEE Access*, *8*, 139367-139386.
29. Malik, H., Bashir, U., & Ahmad, A. (2022). Multi-classification neural network model for detection of abnormal heartbeat audio signals. *Biomedical Engineering Advances*, *4*, 100048.
30. Jabbar, J., Mehmood, H., & Malik, H. (2020). Security of cloud computing: belongings for the generations. *International Journal of Engineering & Technology*, *9*(2), 454-457.
31. Komal, A., & Malik, H. (2022, April). Transfer learning method with deep residual network for COVID-19 diagnosis using chest radiographs images. In *Proceedings of International Conference on Information Technology and Applications: ICITA 2021* (pp. 145-159). Singapore: Springer Nature Singapore.
32. Mahmood, M. A., Malik, H., Khan, A. H., Adnan, M., & Khan, M. I. A. (2022). Neural Network-Based Prediction of Potential Ribonucleic Acid Aptamers to Target Protein. *Journal of Computing & Biomedical Informatics*, *4*(01), 21-36.
33. Malik, H., Chaudhry, M. U., & Jasinski, M. (2022). Deep Learning for Molecular Thermodynamics. *Energies*, *15*(24), 9344.
34. Jabbar, J., Hussain, M., Malik, H., Gani, A., Khan, A. H., & Shiraz, M. (2022). Deep Learning Based Classification of Wrist Cracks from X-ray Imaging. *CMC-COMPUTERS MATERIALS & CONTINUA*, *73*(1), 1827-1844.
35. Khan, M. S. S., Akbar, M. O., Malik, H., Khan, A. H., & Akbar, Z. (2021, November). Variable Generalization Evaluation of Supervised Learning Models for Detection of Spam Messages. In *2021 International Conference on Innovative Computing (ICIC)* (pp. 1-7). IEEE.