APPLICATION OF FEATURE CORRELATION ON MEDICAL BIG DATA ANALYSIS USING DEEP LEARNING

SUBHAM KUMAR^{#1} & DR. FARHA HANEEF^{*2}

^{1#}M.Tech Scholar, Department of CSE, OIST, Bhopal (M.P), India

^{*2}Professor, Department of CSE, OIST, Bhopal (M.P), India

ABSTRACT

This research describes a fully intelligent and automated brain tumor segmentation system depending on Deep Convolutional Neural Networks with feature correlation. Hierarchical CNN model motivated the system design, which was tweaked to improve brain tumor detection performance. Both the BraTS' 2018 training and testing datasets were used to implement and refine the proposed method. Additionally, we have compared the two models i.e; with feature correlation learning model and without feature correlation learning model on the basis of accuracy and model loss and it revealed that with feature correlation learning model achieved the accuracy of 99% with minimum loss value of 2%. This shows that transfer-learning model outperforms without feature correlation learning model.

KEYWORDS: Feature Correlation, Medical Big Data Analysis

Received: Oct 04, 2021; **Accepted:** Oct 24, 2021; **Published:** Oct 30, 2021; **Paper Id:** IJMPERDOCT202192

1. INTRODUCTION

Big data analysis based on the internet of things (IoT) is an intelligent computing method that tries to change conventional objects into intelligent ones using web services, data and analytics, and other methods in today's world. Considering IoT-based big data analysis, Deep Learning (DL) is becoming a gold-standard machine-learning algorithm [1]. Hence, IoT-based DL-based big data analysis has broadened the research area of healthcare monitoring innovation. IoT is one of the most important sources of big data since it concentrates on linking a large variety of smart devices to the Internet, capturing, and reporting the state of their surroundings on a regular basis. Furthermore, medical-based big data has a significant effect on health care decision-making. Some innovative study has been carried to explain the common features of massive data of volume, velocity, and variety. Moreover, in light of "6V" features such as volume, velocity, variety, veracity, variability, and value, generalization of identifying big data across characterizing IoT-based medical big data is discussed. The Internet of Medical Things (IoMT) intends to enhance healthcare systems by collecting information from clinical devices and applications. It uses a variety of sensors to track a patient's status in real times, including a visual sensor, an accelerometer sensor, a carbon dioxide sensor, a temperature probe, electrocardiogram (ECG) or electroencephalogram or electromyogram sensors, a gyroscope sensor, and a blood pressure sensor [2]. Automated medical devices keep track of their patients' fitness, gather patient information, and transmit it to doctors via cloud computing environment. The key difficulty for IoMT is how to govern clinical applications that create a big volume of medical data from linked smart devices. Translation of patient information to viable therapeutics is a key objective of medical artificial intelligence (AI). In biomedicine, machine learning models encounter unique hurdles such as the width of the library needed to train the system, data input translation issues, transmission, over-fitting, and ignorance of covariates, among others [3]. They may necessitate new foundations while rendering previously established

processes outdated. Deep neural network techniques, on the other hand, may have significant advantages.

II. RELATED WORK

Md. Asadur Rahman [1] Many professors, researchers, and commercial entities have been drawn to DL and IoT, and this trend will continue in the future years. With the suggested DL algorithm incorporating CNN and LSTM for health monitoring with improved accuracy, the results are visibly good, making healthcare services more of a pleasurable experience. IoT and deep learning work together to establish a data generation chain that aids in fine-tuning and development in the field. This assessment examines existing approaches, as well as their advances, working procedures, limitations, applications, and future research in the fields of deep learning, IoT, and data mining. Fabian V. Filipp [3] Genomic signatures, image processing, and drug development are just a few of the biomedical fields that have embraced big data and new learning-based tools. Prior to the introduction of many other emerging technologies, artificial intelligence is another crucial point towards that rationalized, data-driven process in the health and medical industries. From conventional methods relying on contributes from natural environment to brute-force inspection utilising robotics, artificial intelligence is yet another pivotal moment toward a rationalized, data-driven procedure in the health and medical industries. Ai-Min Yang, Yang Han [4] Developers gathered 50000 cancer recurrence data in this study, which included liver cancer, lung cancer, renal cancer, breast cancer, uterine cancer, gastric cancer, and colorectal cancer. To begin, they obtain the ratings of each index from each person's health assessment (medical evaluation) and create a cancer recurrence method is proposed based on the D-TSVR algorithm. The increased TSVR accuracy rate is much greater than other models when evaluated to e-TSVR and CNN algorithms. The predicted accuracy percentage is greater than 91%. It demonstrates that the cancer recurring prediction model described in this research is more suitable for forecasting cancer patients' recurrence time and improving their survival rate through clinical interventions. J.-W. Baek, K. Chung [5] Using multiple-regression, this study indicated the Context-DNN model for predicting depression risk. Each DNN model uses context as its input data. Four performance reviews were undertaken with the suggested Context-DNN. Because the DNN has a smaller loss function and precision than the Context-DNN, the multiple regression analysis assessed the variable's evidential support as unit as a consequence of the conformance evaluation. Second, the Context-DNN model was evaluated using cross-validation. Third, the accuracy and recall of the model were compared to that of a general DNN. As a consequence, Context-DNN outperformed regular DNN in terms of functionality. The suggested model taught DNNs individually to increase accuracy and used regression analysis to predictors. Hence, the suggested context-DNN model enables accurate assessment of one's context for depression risk, continued mental health care, and depression prevention. P.-S. Chiu *et al.* [6] The integration of AIoT and mobile voice assistants into campuses will be a modern intelligent campus trend in the development. Researchers developed a smart campus virtual assistant using a deep convolutional neural network (CNN) and a long-short term memory recurrent neural network in this research (RNN-LSTM). Speech recognition, emotion recognition, Chinese characters vector (word embedding), robot conversation, and a campus app are all part of this work. The Unity 3D game engine models the front-end avatar, which includes appropriate body movements based on the user's expression and discourse. In brief phrase emotion detection, this study may obtain a highest accuracy of 95.6 percent. Chatbot can appropriately reply students' queries concerning campus maps, class configuration, nearby restaurants, and fundamental school affairs using guided questions and Chinese Word Embedding, as well as providing basic contextual discourse depending on the recipient's feelings. The Unity 3D model is used to construct AR context conversation robots in this study. Parminder Singh [7] In this study, researchers introduce the concept for heart disease and multi-disease big healthcare data analysis that combines beetles swarm optimization and an adaptive neuro-fuzzy inference

system. In the second study, a revised version of the crow search algorithm (CSA) is used to choose features. For appropriate feature extraction, the data is separated first and then combined. Kaggle was used to collect the cardiovascular disease dataset. The multi-disease data was received from the United States Department of Health and Human Services. The suggested BSO-ANFIS model is compared to the HODBNN, HRFLM, v2 DNN, and ICA(MH) models for cardiovascular disease analysis during the first research study. In the second scenario, the effectiveness of a neural network, a svm, and a traditional ANFIS for multi-disease analysis is determined. Both studies show that the new BSO-ANFIS model outperforms the existing models. KC Santosh [8] Researchers investigated 40 research publications (16 CT + 24 X-ray) other than pre-prints and scientific journals for Covid-19 screening in this report. In this study, they only look at medical imaging technologies to see if their implementation is affected by dataset size. They revealed that the quality did not enhance with the amount of the dataset in both image modalities: CT and X-ray pictures. They have also seen the potential for over-fitting in early 2020. A huge dataset, on the other side, did not appear to improve outcomes because it did not promise that they have all conceivable Covid-19 appearances. Furthermore, they found that data augmentation was effective in enhancing results. However, researchers were unable to determine if the augmentation process can result in the emergence of novel Covid-19 forms. According to reports in the computer vision area, transfer learning may be able to create a Covid-19 deep learning model with less data. Qin Song [9] This study provides a big-data-driven technique that uses DNN to forecast gastrointestinal infection morbidities regarding food contamination statistics and other associated data, and then uses heuristic optimization to generate an emergency medicine procurement process to manage for the future cases. Numerical simulations show that the suggested approach outperforms conventional ways, and in two real case studies, the technique has an average prediction error of 8% and an anticipated recovery rate of 99 percent, which is significantly better than the hospital's traditional approach.

III. METHODOLOGY

When we receive the digital information related to health or medical (electronic health records) EHRs from many sources, we must analyze it. The flowchart of proposed methodology is presented in figure 1. The initial stage is to clean up the information, which includes dealing with missing and erroneous values as well as performing normalization. Unnecessary gaps, characters, or keywords may appear in the featured data. The majority of healthcare data available is unorganized. Because medical data is so different, it is difficult to aggregate or combine it. In practice, many medical and biological databases are partial, comprising a percentage of missing attribute values. To address the issue of insufficient databases, missing value imputation can be used. Because the gathered information often comprises a limited number of image dimensions, feature selection can be used to improve pattern recognition. After feature selection, the process of feature extraction takes place and then the extracted features were categorized into different subsets. These subsets are then sent to rank based chi square, where the comparison between actual results and expected results. This process then leads towards estimation of feature entropy and lastly segregated result is sent to ensemble auto-encoder where it is decided that what the type of disease is and if person is healthy and fit then it is also known to the encoder.

The steps of methodology is as follows:

A. Medical Data

In this step, medical dataset is taken for analysis purpose.

B. Feature Extraction

Feature extraction is the procedure of highlighting relevant data features or qualities. Image classification and discovering similar themes throughout a huge collection of articles are two examples of this method. If the data includes many dimensions (for example, keywords in an article), feature extraction can be used to provide a more compact summary of the process.

Figure 1: Proposed Flowchart

One of the most important processes in computer vision applications is feature selection or feature reduction. A feature extraction extracts the most effective attributes from a feature vector and discards the useless information. This is also a hotly debated field of research in machine learning and pattern recognition. A large number of pictures are analysed for authentication in medical imaging. Images can be at high resolution in certain cases, containing more attributes. Because of the vast amount of data, the feature extraction technique yields a higher dimensionality feature set. Whenever utilizing feature extraction methods, for example, all features contribute to the new extracted features. As a result, one might expect that all (or almost all) material essential for discriminating across data classes is considered and represented in an retrieved feature set. Once the feature selection strategy is applied (for example, forward or backward feature selection), just a subset of features is selected as a new feature set. In this step rank-based chi square algorithm is used for feature selection.

The chi-square test aids in feature selection by studying the connection between the features. A random variable follows chi-square distribution if it can be written as a sum of squared standard normal variables.

$$x^2 = \sum z_i^2 \quad (\text{i})$$

The largest amount of logically independent variables with the freedom to vary is referred to as degrees of freedom. Simply said, it is the total number of data points minus the number of independent constraints imposed on the findings.

$$df = n - 1 \quad (\text{ii})$$

Degree of freedom is denoted as df,

Sample size is abbreviated as n,

The test of chi square can be represented by using below mentioned formula and creating a graph, as shown above. In statistic, the chi-square test was performed to determine the independence of two variables. We can get observed count o and expected count e from the information of two parameters. Chi-Square calculates the difference between the expected count e and the observed count o .

$$x^2 = \sum \frac{(o_i - e_i)^2}{e_i} \quad (\text{iii})$$

When two features are separate, the observed count is close to the expected count, and hence the Chi-Square value is less. A high Chi-Square value suggests that the independent assumption is false. In other words, the greater the Chi-Square value, the more reliant the feature is on the reaction and can be chosen for training phase. The Chi-Square Test is performed in the following steps:

C. Feature entropy estimation

It is a method for economic modelers to use information to improve the hypotheses they makes about parameters in economic models.

D. Training Network

The architecture of the CNN employed in this study is hierarchical. It is made up of a contracting path (on the left) and an extending path (on the right) (right side). Instead of plain blocks, the contracting path comprises of three pre-activated residual blocks. Each block includes two processing units, and each has a Batch Normalization (BN) layer, a Rectified Linear Unit (ReLU) and a convolutional layer but instead of utilizing Max pooling. The quantity of feature streams doubles with every down sampling phase. A fourth residual unit follows the shrinking path, acting as a link between the two trajectories. Similarly, the extending path is made up of three residual blocks. Prior to each block, an up-sampling procedure is performed, which raises the feature map size by a factor of two. The feature maps belonging to the shrinking path are then concatenated after a 2×2 convolution. Finding the network properties (weights and biases) that minimize a loss function is needed to be focused. This is accomplished in this study by employing the Stochastic Gradient Descent

algorithm (SGD) [10], which changes the parameters in the inverse direction of the gradients at every iteration. To solve the class imbalance issue in brain tumor data, we employed a loss function as eqn 1:

$$E_{weighted\ cross} = -\frac{1}{K} \sum_k \sum_i^L w_i G_{ik} \log(P_{ik}) \quad (1)$$

Here, L stands for the overall number of labels, while K stands for the batch size. The weight allocated to the i^{th} label. P_{ik} and G_{ik} denoting the valuation of the (i^{th} , k^{th}) pixel of the binary picture and the binary quantitative measurements image, respectively.

E. Brain Tumoral Structures Prediction

Forecasting can be done after the network has been trained. The network must be able to produce a segregated image after receiving the four MRI modes of an elongated volume that it has neither evaluated or observed previously.

IV. RESULT ANALYSIS

In this research, we used pre-operative heterogeneous MRI scans from both the mentoring and validating databases of the BraTS'2018 obstacle to test and assess. Our investigation is based on the comparative study of the two models namely, with feature correlation learning and Without feature correlation learning. After performing the experiments using both the abovementioned models the outcome revealed in the terms of accuracy of model and model loss within particular epochs. We are going to describe in detail about the particular model with its accuracy and loss. Model accuracy in without feature correlation model- The graph presented below demonstrates that initially when the training of the model is started, the accuracy increases and suddenly drops as depicted. The increase in value of epochs the accuracy of the model is increases but this graph is not constant, ups and down in the accuracy continues. As the average accuracy rate can be given as 75%.

Figure 2: Model Accuracy in without Feature Correlation Model

Model loss in without feature correlation learning model- Loss of the model during training phase is shown in figure below. The graph depicts that initially as the training of the model takes place the loss value is at its peak but as the epoch value increases gradually, the loss value minimizes.

Figure 3: Loss in without Feature Correlation Model

Model accuracy in with feature correlation learning model- The graph given below demonstrates the accuracy of model, which gradually raises with the increase in the epoch values. The figure shows that the graph is not constantly increases. At certain values the accuracy drops and then raises, the average accuracy rate calculated as 98%.

Figure 4: Accuracy of the Transfer Learning Model

Model loss in with feature correlation learning model- The loss in this model reduces as the increase in the value of epoch. The value of loss is at peak in the initial phase of training but then the value decreases as the increment in the time duration.

Figure 5: Loss of the Transfer Learning Model

The comparative graph of both the models in the terms of the accuracy and loss is shown below. If we discuss about accuracy then with feature correlation learning model performs much better with the accuracy rate of 97% and in terms of loss it is stated that the value of loss in without feature correlation learning model is high as compared to loss in with feature correlation learning model.

Table 1: Comparative Analysis

Model	Accuracy	Loss
With Feature Correlation Learning Model	98%	2%
Without Feature Correlation Learning Model	75%	35%

The graphical analysis of the both the models is shown in figure 6. The research is evaluated on the basis of the two learning modeling techniques: with feature correlation learning model and without feature correlation learning model. All-round, the approach with feature correlation learning model leads the model without feature correlation learning model in terminology of accuracy and loss, according to the results.

Figure 6: Graphical Analysis of both the models with accuracy and loss values

V. CONCLUSIONS

A fully autonomous and precise approach for identifying entire brain tumors using a deep convolutional network depending on hierarchical medical imaging structure is considered for the analysis. Furthermore, we analyzed the accuracy and model loss of the two learning models, namely the with feature correlation learning model and without feature correlation learning model. The performance of the above-mentioned comparative models was evaluated, and it was discovered that the with feature correlation learning model model attained a 98% accuracy with a minimal loss value of 2%. And the accuracy of without feature correlation learning model was evaluated as 75% and the model loss is 35%. This demonstrates that a with feature correlation learning model outperforms a model that does not use feature correlation.

REFERENCES

1. Md. Asadur Rahman. *A Narrative Analysis on Deep Learning in IoT based Medical Big Data Analysis with Future Perspectives*. Access International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 7-9, 2019.
2. Salome Oniani. *Artificial Intelligence for Internet of Things and Enhanced Medical Systems*. Access, Springer, 2021, https://doi.org/10.1007/978-981-15-5495-7_3
3. Fabian V. Filipp. *Opportunities for Artificial Intelligence in advancing precision Medicine*. Access Springer, vol. 7, pp. 208-213, 2019, <https://doi.org/10.1007/s40142-019-00177-4>
4. Ai-Min Yang. *D-TSVR Recurrence Prediction Driven by Medical Big Data in Cancer*, Access IEEE, 2020, DOI 10.1109/TII.2020.3011675.
5. JI-WON BAEK. *Context Deep Neural Network Model for Predicting Depression Risk Using Multiple Regression*, Access IEEE, vol. 8, 2020, Digital Object Identifier 10.1109/ACCESS.2020.2968393

6. P.-S. Chiu. *Enabling Intelligent Environment by the Design of Emotionally Aware Virtual Assistant*, Access IEEE, vol. 8, 2020, Digital Object Identifier 10.1109/ACCESS.2020.2984383.
7. Parminder Singh. *Multi-disease big data analysis using beetle swarm optimization and an adaptive neuro-fuzzy inference system*, Access Springer, 2021, <https://doi.org/10.1007/s00521-021-05798-x>
8. KC Santosh. *Covid-19 Imaging Tools: How Big Data is Big?* Access Springer, 2021, <https://doi.org/10.1007/s10916-021-01747-2>
9. Qin Song. *Emergency Drug Procurement Planning Based on Big-Data Driven Morbidity Prediction*, Access IEEE, 2018, DOI 10.1109/TII.2018.2870879
10. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press, Cambridge. 2016.