Multimodal Diagnosis for Pulmonary Embolism from EHR Data and CT Images

Zhuo ZHI[1], Moe Elbadawi[2], Adam Daneshmend[3], Mine Orlu[2], Abdul Basit[2], Andreas Demosthenous[1], and Miguel Rodrigues[1]

***Abstract*— Pulmonary Embolism (PE) is a severe medical condition that can pose a significant risk to life. Traditional deep learning methods for PE diagnosis are based on Computed Tomography (CT) images and do not consider the patient's clinical context. To make full use of patient's clinical information, this article presents a multimodal fusion model ingesting Electronic Health Record (EHR) data and CT images for PE diagnosis. The proposed model is based on multilayer perception and convolutional neural networks. To remove the invalid information in the EHR data, the multidimensional scaling algorithm is performed for feature dimension reduction. The EHR data and CT images of 600 patients are used for experiments. The experiment results show that the proposed models outperform existing methods and the multimodal fusion model shows better performance than the single-input model.**

I. INTRODUCTION

The incidence of Pulmonary Embolism (PE) has been increasing in recent years. It brings an age-standardized mortality rate ranging from 0 to 24 deaths per 100 000 population-years [1]. The standard method for PE diagnosis involves visual inspection of the Computed Tomography (CT) images [2]. However, with limited healthcare resources and expensive specialist fees, delays and misdiagnoses of PE are common [3]. There remains space for improvement of PEimages diagnosis based on CT imaging.

With the development of Deep Learning (DL) technology, many models have been built for PE diagnosis based on CT, showcasing satisfactory results [4], [5], [6]. Although PE diagnosis by DL leads to reasonable results, it still differs from radiologists' clinical, which mainly manifests in the use of the patient's clinical context. To be specific, doctors leverage lab test results, prior diagnosis and disease history in the Electronic Health Record (EHR) to diagnose PE more accurately. According to [7], 83% of CT reports become more accurate when considering full clinical information. Therefore, combining EHR data and CT images can be a promising avenue.

Multimodal learning provides a framework to combine different kinds of medical information for DL based disease diagnosis. To achieve early detection of Alzheimer's disease stage, J. Venugopalan et al. [8] propose a multimodal DL model to analyze images and clinical data integrally, showing it outperforms single data modality based models. For diabetic kidney disease, the multimodal model based on EHR data and plasma biomarkers achieves higher prediction accuracy than the single-input model [9]. See also [10], [11]. However, using multimodal models for PE diagnosis brings about new challenges due to the 3D nature of CT images.

Multimodal learning has shown satisfying performance in the medical field, however, there is still much work to be carried out to apply it in PE diagnosis due to limited open dataset and the 3D particularity of CT images.

Recently, a team published RadFusion: a multimodal dataset that includes the EHR data and CT images labeled for pulmonary embolisms [12]. The team also utilizes 3D Convolutional Neural Network (CNN) and ElasticNet to build the multimodal fusion model for PE diagnosis. However, the method shows some deficiencies. For example, all features in the EHR data are involved in the model. In fact, features that contribute to PE diagnosis are only part of them. The model's performance could be affected if unnecessary features are introduced. In addition, the 3D CNN model requires that the number of 2D slices in each 3D CT image (equal to the depth of 3D images) needs to be the same. In fact, the numbers show huge difference. Therefore, the depth of 3D CT images is compressed to meet the 3D CNN model requirements, resulting in a decrease in sample diversity.

Hence, this article proposes a multimodal fusion model based on Multilayer Perceptron (MLP) and 2D CNN for PE diagnosis. MLP is commonly utilized for classification problems with large-scale data due to its advantages like high degree of parallelism, highly nonlinear global action, excellent fault tolerance ability and associative memory function [13]. The depth of the original 3D CT image will be retained completely in 2D CNN model. Moreover, the Multidimensional Scaling (MDS) algorithm is applied to reduce the dimension of EHR data in order to keep important features. The MDS algorithm does not need prior knowledge and has high computational efficiency. Comparison experiments on 600 patients' data from RadFusion are implemented. Results show that the proposed method could achieve satisfying performance. The contribution of this study can be summarized as follows:

1) A high-performance multimodal fusion model based on MLP-2D CNN is proposed for PE diagnosis.
2) The MDS algorithm is applied to reduce the data dimension and improve the over-fitting phenomenon.
3) The 2D CNN model is designed to address the issue that the depth of 3D CT image is compressed in the 3D CNN model.

*This work was supported by Engineering and Physical Sciences Research Council (EPSRC), UK

[1]Department of Electronic and Electrical Engineering, University College London, London, UK

[2]UCL School of Pharmacy, University College London, London, UK

[3]Imperial College Healthcare NHS Trust, London, UK

The rest of this article is organized as follows. Section II introduces the proposed method. Section III presents the experiment and results analysis. Section IV draws the conclusion.

II. The Proposed Method

In this section, the proposed EHR-only model, image-only model and multimodal fusion model are introduced firstly. Then, the MDS algorithm is also explained.

A. The proposed models

Firstly, an EHR-only model based on MLP is designed for PE diagnosis using only EHR data. There are 2912 features for each patient in the EHR data. To remove unimportant features, the MDS algorithm is applied to the original data. Then, to solve the problem that the depth of 3D image will be compressed in the 3D CNN model, an image-only model based on 2D CNN is proposed. Finally, these two models are combined to build the multimodal fusion model. The structure of the proposed models is shown in Fig. 1.

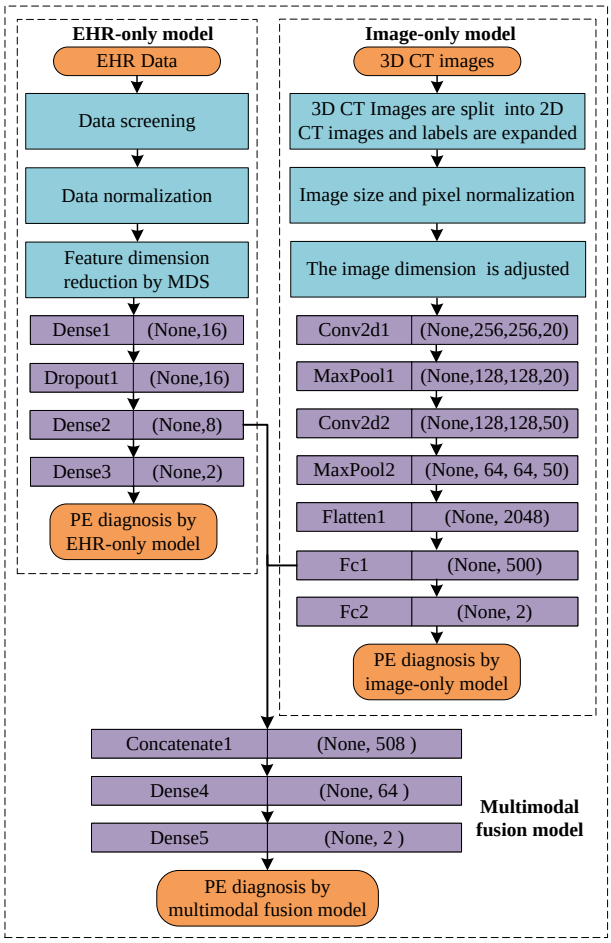


Fig. 1. The structure of the proposed models.

It can be seen from Fig. 1 that the EHR-only model and the image-only model are utilized for processing two kinds of input firstly. Then, the penultimate layers of two models are stitched together and input into a new model to build the multimodal fusion model. These three kinds of models are introduced next.

1) The EHR-only model: The flow chart of EHR-only model is shown in the dotted box on the upper left of Fig. 1. The first step is to screen original EHR data. The empty columns and outliers are removed. Then, to improve the convergence rate of the model, all the data are normalized and scaled to fit the interval [0, 1]. After that, the MDS algorithm is used to reduce the feature dimension. Based on the filtered data, the MLP classification model is built for PE diagnosis. The MLP model utilizes two hidden layers for features and logical processing and one dropout layer for reducing the over-fitting problem. The MLP method is described in [13].

2) The image-only model: The flow chart of the image-only model is shown in the dotted box on the upper right of Fig. 1. The CT image for each patient is presented in the 3D format, which consists of N 2D CT slices (N ranges from 28 to 2600). The depth, width and length of 3D images need to be unified for a 3D CNN, which means that N will be changed and the integrity of the samples is compromised. To solve this problem, a 2D CNN model is designed. In this model, the input is 395872 2D CT slices instead of 1112 3D CT images. Hence, the first step is to split 3D CT Images into 2D CT images and expand labels. Then, the image size and pixel normalization are carried out. For the 2D CNN model, it only needs to unify the width and length for all images. All pixels are scaled to fit the interval [0, 1] for normalization, which can accelerate the model convergence rate. After that, the dimension of all images is adjusted to facilitate the convolution operation. Finally, the classification model based on 2D CNN is built and trained for PE diagnosis. The 2D CNN model utilizes two convolution layers for feature extraction and two pooling layers for feature dimension reduction.

3) The multimodal fusion model: To make full use of patients' clinical information, the multimodal fusion model is built by combining EHR-only model and image-only model together. The last layers of the EHR-only model and the image-only model are used to map model tensors to the labels. In the multimodal fusion model, the last layers of them are skipped and the penultimate layers of two models are stitched together to create a new tensor. The new tensor is fed into subsequent neural network layers for classification, which realizes the prediction by multimodal fusion information.

B. The MDS algorithm

The MDS algorithm aims to map the high-dimensional data to the low-dimensional data on the premise that the distance between the samples in original space and new space is consistent. Suppose there are m samples with d dimensions and the target dimension is d'.

The original sample space is expressed as

$$T = \{x_1, x_2, \cdots, x_m\} \quad , x_i \in R^d. \tag{1}$$

Let matrix D represents the distance between original samples. $dist_{ij}$ is the element in D and refers to the distance between x_i and x_j. The goal of the MDS algorithm is to get the new samples $Z \in R^{d' \times m}, d' \leq d$. The distance between the two samples remains unchanged after mapping, which means $\|z_i - z_j\| = dist_{ij}$. Let $B = Z^T Z \in R^{m \times m}$, where B is the inner product matrix of new samples and $b_{ij} = z_i^T z_j$. The relationship between $dist_{ij}$ and B is

$$dist_{ij}^2 = \|z_i\|^2 + \|z_j\|^2 - 2z_i^T z_j = b_{ii} + bjj - 2b_{ij}. \quad (2)$$

Z is commonly centralized so that $\sum_{i=1}^{m} z_i = 0$ and $\sum_{i=1}^{m} b_{ij} = \sum_{j=1}^{m} b_{ij} = 0$.

Let $dist_{i.}^2 = \frac{1}{m}\sum_{j=1}^{m} dist_{ij}^2$, $dist_{.j}^2 = \frac{1}{m}\sum_{i=1}^{m} dist_{ij}^2$ and $dist_{..}^2 = \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} dist_{ij}^2$, the element in B can be expressed as

$$b_{ij} = -\frac{1}{2}\left(dist_{ij}^2 - dist_{i.}^2 - dist_{.j}^2 + dist_{..}^2\right). \quad (3)$$

Since B is a symmetric matrix, the eigendecomposition of B can be obtained by

$$B = V \Lambda V^T, \quad (4)$$

where $\Lambda = \mathrm{diag}\,(\lambda_1, \lambda_2, \cdots, \lambda_d)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ and V is the corresponding eigenvector matrix. Suppose there are d^* nonzero eigenvalues and V_* is the eigenvector matrix, the objective Z is expressed as

$$Z = \Lambda_*^{\frac{1}{2}} V_*^T \in R^{d^* \times m}. \quad (5)$$

III. Experiments

Three kinds of experiments are implemented to evaluate the performance of proposed models. The comparative methods and the dataset are derived from the novel study [12].

1) For the EHR-only model, we first compared different values of the target dimension of the MDS algorithm. Then, we evaluate the performance of the proposed MLP model with the ElasticNet model.
2) For the image-only model, we evaluate the performance of the proposed 2D CNN model with the 3D CNN model.
3) For the multimodal fusion model, we evaluate the performance of the proposed MLP-2D CNN model with the ElasticNet-3D CNN model.

The software and hardware platforms involved in experiments are shown in Table I.

TABLE I
The Experiment Environment.

Software/ hardware	Model/version
Operation system	CentOS 8.0
GPU	NVIDIA Tesla A100
CPU	AMD EPYC 7543
Model framework	Tensorflow 2.4

A. The dataset

The dataset consists of three parts: the EHR data, CT images and the label. The content and structure of the data is described next.

1) The EHR data: The EHR dataset is presented as a list. The vertical index of the list is the identification code of patients. The horizontal indices of the list are 2912 features of patients. These features include

- Demographic features
  Demographic features consist of one-hot encoded gender, race, smoking habits and age.
- Vitals
  Vitals include systolic and diastolic blood pressure, height, weight, body mass index, temperature, respiration rate, pulse oximetry and heart rate.
- Inpatient and outpatient medications
  641 unique classes of drugs are identified for inpatient and outpatient medication. For each medication, the intake frequency and a label of whether the patient takes it are presented.
- Diagnosis code
  141 unique diagnosis groupings are given. For each grouping, there is a binary label indicating the presence/ absence of the corresponding condition.
- Laboratory tests
  22 categories are generated for all laboratory tests. Similarly, a binary label is used to present the presence/ absence of the test.

2) The CT images: The CT images are presented as separate files with .npy format. For each patient, the CT images data consist of N 2D-CT slices (N ranges from 28 to 2600). Each slice has 1 channel with the gray color type and the size is 512*512 pixels. The total number of all slices is 395872.

3) The label: The label is presented as a list. For each patient, the label is given as 0 or 1. 0 refers to negative PE and 1 indicates positive PE. All labels are generated by manual review by three board-certified radiologists.

B. Experiments with EHR-only model

As mentioned before, the original EHR data contain 2912 features. After screening the data, 1286 features are selected as the input of the EHR-only model. The MDS algorithm is performed then. During this process, the value of the target dimension of the MDS algorithm is set to be 1000, 500, 200, 100, 50, 20, 10, 5 and 2 for the experiment. 600 patients are selected for the experiment and they are split into three parts randomly: 120 patients to be the test data, 384 patients to be the train data and 96 patients to be the validation data. The evaluation indices are the accuracy, F1 score and the training time. Among these indices, accuracy refers to the proportion of the samples that receive the correct prediction. F1 score is the harmonic mean of precision and recall rate. F1 score is positively correlated with the model performance. The experimental results are shown in Table II.

It can be seen from Table II that if all the features are input into the model, it only achieves the accuracy of 52%, which

TABLE II

THE EXPERIMENT RESULT OF THE MDS ALGORITHM.

No.	The value of target dimension	Accuracy	F1 score	Computing time / min
1	Without MDS	52.2%	0.494	10.2
2	1000	54.1%	0.513	9.1
3	500	54.4%	0.526	8.3
4	200	58.3%	0.541	6.9
5	100	57.7%	0.543	6.0
6	50	62.3%	0.614	5.5
7	20	66.5%	0.597	5.1
8	10	66.8%	0.647	4.4
9	5	69.2%	0.665	4.1
10	2	66.7%	0.634	4.0

is approximate to the random guess. When the dimension of EHR data is continuously reduced by the MDS algorithm, the accuracy and F1 score keep increasing and they reach the maximum values when the target dimension is 5. The accuracy is increased by 27.8% and the F1 score is increased by 28.3%. Moreover, the training time also reduces gradually. Therefore, the target dimension of the MDS algorithm will be set as 5 in the subsequent experiments. The training process of these experiments are also recorded and the training history of No. 1 experiment and No. 10 experiment is shown in Fig. 2.

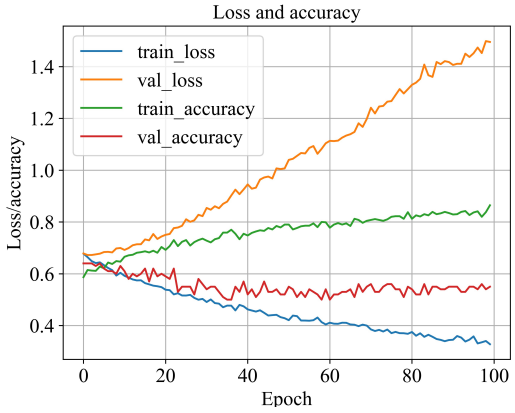
It can be seen from Fig. 2 (a) that the val_loss keeps increasing and the train_loss shows the opposite trend when all the features are involved in the training process, which means the serious over-fitting phenomenon happens. The problem is solved by removing unimportant features and the val_loss and train_loss present a normal trend in Fig. 2 (b). After determining the target dimension value of the MDS algorithm, the experiment for comparing the performance of the ElasticNet with the MLP model is implemented. The experiment result is shown in Table III.

TABLE III

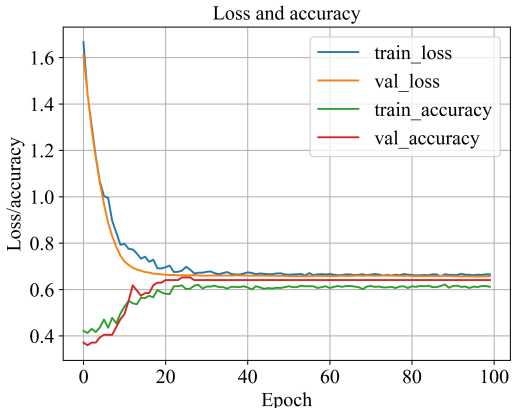
THE EXPERIMENT RESULT OF EHR-ONLY MODELS.

Model name	Accuracy	F1 score	Training time / min
ElasticNet	66.6%	0.652	7.2
MLP	69.2%	0.695	4.1

It can be seen from Table III that the proposed MLP model achieves the higher accuracy and F1 score compared to the ElasticNet model. Specifically, the accuracy is increased by 4.0% and the F1 score is increased by 6.6%. What's more, the training time is reduced significantly by 75.6%. In summary, the feature dimension reduction is vital in dealing with EHR data and can significantly improve the over-fitting phenomenon as well as the computational efficiency. Although the proposed EHR-only model based on MLP has higher accuracy and F1 score than the existing method, it can not be used alone for PE diagnosis.



(a)



(b)

Fig. 2. The training record of No. 1 experiment and No. 10 experiment. (a) The training record of No. 1 experiment. (b) The training record of No. 10 experiment.

C. Experiment with image-only model

The 2D CNN model and the 3D CNN model are compared in this experiment. The evaluation indices are accuracy, F1 score and the training time. For the 3D CNN model, the size of each patient's 3D image is unified to 356*256*256 (depth*width*length) to minimize the loss of sample diversity (356 is the average of all 2D slices). The image size for the 2D CNN model is set to be 256*256 (width*length). The data partitioning method is the same as before: 120 patients to be the test data, 384 patients to be the train data and 96 patients to be the validation data. The experimental results are shown in Table IV.

TABLE IV

THE EXPERIMENT RESULT OF IMAGE-ONLY MODELS.

Model name	Accuracy	F1 score	Training time / min
3D CNN	78.6%	0.802	275.1
2D CNN	95.3%	0.948	95.7

It can be seen in Table IV that the 2D CNN model can achieve higher accuracy, F1 score than the 3D CNN model.

Specifically, the accuracy is increased by 21.2% and the F1 score is increased by 18.2%. This phenomenon may be due to that the diversity of 3D CNN model samples is affected by unifying depth. In addition, the process of unifying depth requires complex calculation, which results in much longer training time.

D. The experiment of the multimodal fusion model

In this experiment, the performance of the MLP-2D CNN model and the ElasticNet-3D CNN model is compared. The evaluation indices are accuracy, F1 score and the training time. The data are divided as same as before: 120 patients are the test data, 384 patients are the train data and 96 patients are the validation data. The experimental results are shown in Table V.

TABLE V
THE EXPERIMENT RESULT OF IMAGE-ONLY MODELS.

Model name	Accuracy	F1 score	Training time / min
ElasticNet-3D CNN	92.4%	0.897	320.2
MLP-2D CNN	97.3%	0.964	101.6

It can be seen from Table V that the accuracy and the F1 score of the MLP-2D CNN model has different degrees of improvement than the ElasticNet-3D CNN model, 5.3% and 7.5%. In addition, the computational efficiency of the MLP-2D CNN is significantly improved. The time consumption is 215.2% lower than the ElasticNet-3D CNN model. To get a more intuitive comparison, the indices of all models are shown in Table VI.

TABLE VI
THE INDICES OF ALL MODELS.

Model type	Model name	Accuracy	F1 score	Training time / min
EHR-only	MLP	69.2%	0.695	4.1
EHR-only	ElasticNet	66.6%	0.652	7.2
Image-only	2D CNN	95.3%	0.948	95.7
Image-only	3D CNN	78.6%	0.802	275.1
Multimodal fusion	MLP-2D CNN	97.3%	0.964	101.6
Multimodal fusion	ElasticNet-3D CNN	92.4%	0.897	320.2

It can also be seen from Table VI that the proposed MLP-2D CNN model achieves the highest accuracy and F1 score among three kinds of models. In addition, the two multimodal fusion models both have the higher accuracy and F1 score than their constituent models.

IV. CONCLUSION

To combine the EHR data and CT images for PE diagnosis, this article designs three kinds of model: the EHR-Only model based on MLP, the image-Only model based on 2D CNN and the multimodal model based on MLP-2D CNN. The MDS algorithm is applied to remove redundant features in EHR data. EHR data and CT images of 600 patients are involved to evaluate the proposed models. At the same time, the ElasticNet model, the 3D CNN model and the ElasticNet-3D CNN model are selected for comparison.

The experiment results show that the MDS algorithm can effectively solve the over-fitting phenomenon. The proposed models have better performance than existing methods in accuracy, F1 score and computational efficiency. The EHR-Only model, the image-only model and the multimodal fusion model achieve the accuracy at 69.2%, 95.3% and 97.3%, respectively. Moreover, all the multimodal fusion models show better performance than corresponding single-input models. To conclude, combining EHR data and CT images can be an effective method for PE diagnosis.

ACKNOWLEDGMENT

The Institution's Ethical Review Board approved all experimental procedures involving human subjects.

REFERENCES

[1] Stefano Barco, Luca Valerio, et al. Global reporting of pulmonary embolism–related deaths in the world health organization mortality database: Vital registration data from 123 countries. *Research and Practice in Thrombosis and Haemostasis*, 5(5):e12520, 2021.

[2] Shih-Cheng Huang, Pareek, et al. Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific reports*, 10(1):1–9, 2020.

[3] José Luis Alonso-Martínez, Sánchez, et al. Delay and misdiagnosis in sub-massive and non-massive acute pulmonary embolism. *European journal of internal medicine*, 21(4):278–282, 2010.

[4] Martine Remy-Jardin, Faivre, et al. Machine learning and deep neural network applications in the thorax: pulmonary embolism, chronic thromboembolic pulmonary hypertension, aorta, and chronic obstructive pulmonary disease. *Journal of thoracic imaging*, 35:S40–S48, 2020.

[5] Errol Colak, Kitamura, et al. The rsna pulmonary embolism ct dataset. *Radiology: Artificial Intelligence*, 3(2):e200254, 2021.

[6] Shelly Soffer, Eyal Klang, et al. Deep learning for pulmonary embolism detection on computed tomography pulmonary angiogram: a systematic review and meta-analysis. *Scientific reports*, 11(1):1–8, 2021.

[7] Adones Leslie, AJ Jones, and PR Goddard. The influence of clinical information on the reporting of ct by radiologists. *The British journal of radiology*, 73(874):1052–1055, 2000.

[8] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):1–13, 2021.

[9] Lili Chan, Girish N Nadkarni, et al. Derivation and validation of a machine learning risk score using biomarker and electronic patient data to predict progression of diabetic kidney disease. *Diabetologia*, pages 1–12, 2021.

[10] Shikha Purwar, Rajiv Kumar Tripathi, et al. Detection of microcytic hypochromia using cbc and blood film features extracted from convolution neural network by different classifiers. *Multimedia Tools and Applications*, 79(7):4573–4595, 2020.

[11] Shih-Cheng Huang, Anuj Pareek, et al. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):1–9, 2020.

[12] Yuyin Zhou, Shih-Cheng Huang, et al. Radfusion: Benchmarking performance and fairness for multimodal pulmonary embolism detection from ct and ehr. *arXiv preprint arXiv:2111.11665*, 2021.

[13] Oludare Isaac Abiodun, Aman Jantan, et al. Comprehensive review of artificial neural network applications to pattern recognition. *IEEE Access*, 7:158820–158846, 2019.