

Deep Multi-Modal Fusion of Clinical and Non-Clinical Data using Early Submission for Enhanced Kidney Disease Prediction

Suhasini Tatiparti¹, Tatiparti B Prasad Reddy²

School of Business, S R University, Warangal, Telangana, India¹

Dept. of CS & IT, Alliance University, Bengaluru, India²

Corresponding author: Suhasini Tatiparti, Email: suhasini.tatiparti@sru.edu.in

Due to the shortage of nephrologists and the scarcity of diagnostic labs in rural areas, chronic kidney disease (CKD) has recently emerged as a major health problem. In order to detect illness from clinical and radiological pictures, automated diagnostic models are necessary. Previous research on CKD diagnosis has mostly focused on the independent use of clinical data and CT scans to create AI algorithms. Incorporating clinical data with Computed Tomography (CT) images, this research seeks to create a Deep Multimodal Fusion approach with a late fusion mechanism for diagnosis. Clinical data were used in the tests with CT scans of the kidney to ensure accuracy. Clinical data and image-extracted features are kept separate in the proposed model by a process termed late fusion. The model scored 99 accuracy, 98.6 recall, 97.8 precision, and 98.2 F-score, showing that combining clinical data with the CTD increases diagnosis accuracy to that of a human expert. An additional method of verification was comparing the proposed system's results with those of a human expert. In addition, the findings validate the feasibility of the proposed approach as a diagnostic aid for chronic kidney disease (CKD) for medical professionals.

Keywords: Chronic kidney disease, CT, Precision, Accuracy, Precision, F-score

1. Introduction

The term "Internet of Things" is used to describe a network of interconnected devices that share data with each other and all of the system's programs. When it comes to the healthcare business and patients, the Internet of Medical Things (IoMT) becomes a specialized, extended branch of IoT that incorporates all connected devices used to provide timely assistance [1]. Communication between devices may occur wirelessly or through wires. The ability to go about daily life while yet receiving constant health monitoring is one of the biggest benefits of IoMT-related remote health monitoring [2]. The size of the modules attached to the body and the need for regular battery replacement or charging made the conventional remote monitoring techniques look cumbersome to the patients [3]. By developing small, ultra-low-power sensor devices and lightweight gearbox protocols, the IoMT revolution is able to address the aforementioned issues and provide viable solutions. Portable patient monitoring units (PPMUs) in ambulances and residences are used for remote health monitoring [4]. A decision support system is used in hospitals for real-time monitoring.

Combining data or information from several sources with different formats and structures is known as "data fusion" [5]. Using the human brain's perceptual system as an example, the notion of data fusion becomes clearer. Illustration of the human brain's perceptual system The human brain integrates data from the five senses (sight, sound, smell, taste, and touch) with the findings taken from the memories of comparable events to produce a coherent picture of the world around it. The application of data fusion principles to the human body demonstrates the significance of this idea. Not only have geospatial systems, defense systems, and intelligence services (for a review, see [7]) found the use of the concept of data fusion and its associated techniques, but so have healthcare settings, where they provide a basis for developing advanced decision support and smart patient monitoring systems [6]. Both the combination of data (COD) and the combination of interpretation (COI) are common methods for achieving data fusion [8]. Both methods are explained in more detail below.

First, features from all available data sources are combined and used to construct a single decision model (i.e., to train a single classifier), as illustrated in Figure 1.

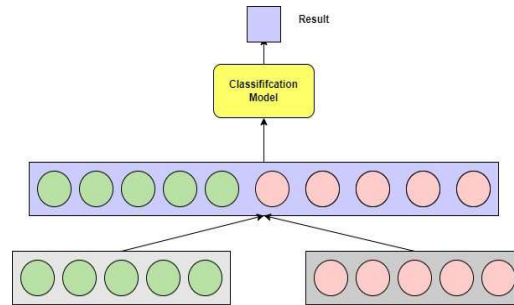


Figure 1 : Combination of Data (features) Model

The second method, called a combination of interpretation (COI), uses many data sources to construct independent decision models, the results of which are then combined by a combiner (also known as a meta-decision model or metaclassifier) to provide a single output. Building ensembles of classifiers (and in particular the stacking approach) [9] is conceptually related to COI. The basic concept of the COD model is shown in Figure 2. The curse of dimensionality plagues COD, while COI is shown to be sub-optimal since it cannot maintain relationships between data from different sources, although both have their flaws. In order to fix these issues, Lee et al. [10] created a general fusion framework (GFF) in which COI and COD are seen as two ends of a continuum.

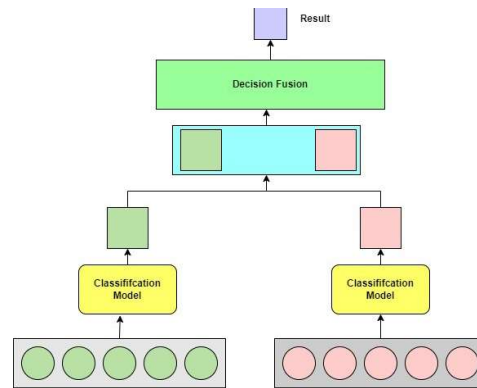


Figure 2 : Combination of Interpretation

The curse of dimensionality plagues COD, while COI is shown to be sub-optimal since it cannot maintain relationships between data from different sources, although both have their flaws. In order to fix these issues, Lee et al. [10] created a general fusion framework (GFF) in which COI and COD are seen as two ends of a continuum.

Several encouraging investigations have been conducted on the topic of using deep learning models to automate diagnosis in medical imaging [7–10]. Previous work has shown the feasibility of effective automated image analysis based on imaging data alone, in contrast to normal clinical practice, which depends on the interpretation of medical imaging in combination with relevant clinical data to guide a suitable diagnosis. Accessing relevant background information at the moment of image interpretation is critical in radiology since making an accurate medical diagnosis on imaging typically relies on pre-test probability, previous diagnosis, clinical and laboratory data, and previous imaging. Eighty-five percent of radiologists think that clinical context is critical for understanding imaging studies [14]. The use of multimodal data fusion for automated clinical outcome prediction and diagnosis has increased during the last three years.

2. Related Work

The literature discusses the use of COD and COI methods from [5,8-11]. In [8], Lanckriet et al. provided an illustration of the COD technique by using a support vector machine (SVM) to predict the function of yeast proteins. Their plan involves using a collection of kernels to merge information on proteins, their constituent amino acid sequences, and the genes they encode. In [9], K. Kourou et al. employed machine learning methods used in cancer diagnosis and prediction to create a variety of classification models informed by the COD strategy. The COI method, as published by Jesneck et al. in [10], combines radiologist interpretation with objective evidence derived from mammograms and patient history to arrive at a breast cancer diagnosis. Classifiers were built using techniques from detection theory. To be more precise, a likelihood ratio-based binary classifier was built for each feature (resulting in multiple classifiers for a given data source), and their combined results were expressed as a joint likelihood ratio of the set of decision variables. Finally, a genetic algorithm was used to optimize the thresholds used in both classification tiers. The notion of ensemble classifiers is discussed in [5], and it is analogous to COI (in particular the stacking scheme) in that it involves the construction of many classifiers. The author explains how to build ensemble classifiers and goes into depth on how to combine the results of individual classifiers. Decision trees, SVMs, and neural networks, including their ensemble, were used by G. Zorluoglu et al. [11] to develop a model for diagnosing breast cancer.

Classifiers' individual performances were compared to those of the ensemble classifier, and the latter was shown to be more accurate.

In [4], the author evaluates the COI and COD by building models for four distinct applications in biomedical image processing. Atlas-based image segmentation, average image tissue segmentation, multi-spectral classification, and deformation-based group morphometry are only some of the tasks that may be performed. Using their adaptability and capacity to provide consistent results, the author compared the performance of COI and COD utilized in the creation of models for the aforementioned functions.

Image data (i.e., pathological pictures, radiological images, and camera photos) and non-image data (i.e., lab test results and clinical data) may be generated from routine clinical visits of a single patient. To better assist diverse clinical choices (such as illness diagnosis and prognosis [1] [2] [3]), the heterogeneous data would give varying perspectives on the same patient. Such methods of reaching decisions, however, might be subjective and qualitative, with significant inter-subject variability [4] [5]. Recent years have seen a proliferation of deep learning-based solutions for multi-modal learning in medical applications, thanks in large part to the fast development of AI technology. When it comes to extracting and modeling the complicated interactions between multiple modalities and outcomes, multi-modal fusion often benefits from deep learning because of the high-level abstraction of complex phenomena inside high-dimensional data [6] [7]. Utilising deep learning techniques, several works have successfully diagnosed or prognosed utilizing a single modality [8, 9], [10]. Since many clinical modalities may have varying information (complementary information of a person) and have varying data formats, fusing the multi-modal data successfully is not a straightforward challenge in method creation.

Diseases are diagnosed and prognoses are made using the patient's multi-modal data (image and non-image data). Radiological images, pathological images, and camera images are the several types of images available[13]. varied imaging data may have varied dimensions (2D, 3D, 4D, etc.), and may be further categorized as pixel-aligned (can be spatially registered and overlaid) or pixel-not-aligned (the pixels in distinct pictures do not have a spatial connection). Lab test results, such as tabular data of demographic traits or free text in lab test reports, are examples of non-image data; other examples include structured genetic sequences and blood test results[16]. Multi-modal learning is a class of methods in machine learning that faces significant difficulties due to the variety of such image and non-image input. For instance, whereas 3D CT/MRI radiological scans give macro-level and geographical information about a tumor, 2D pathology images reveal micro-level morphology[12]. Clinical decision-making involves the structured DNA and mRNA sequences, as well as the clinical data and laboratory test findings, which show the molecular, biological, and chemical properties[17]. In addition, picture data is often bigger and denser (e.g., millions of pixels), whereas non-image data is sparser and lower dimensional. In this context, different types of information necessitate fusion methods that are able to effectively capture the shared and complementary information for rendering better diagnosis and prognosis, and the heterogeneous formats necessitate different preprocessing and feature extraction methods (e.g., different dimensions, images, free text, and tabular data)[15].

3. Data

First, we provide a brief overview of the data collection that was analyzed in this research. The assessment matrix and methods employed will be discussed thereafter. DVR Diagnosis in Hyderabad provided the data for this analysis. In order to accurately diagnose kidney illness, it is necessary to collect data from a wide variety of sources. The sections that follow will detail the various types of information that were collected for this study.

3.1 Clinical Data

The data set contains the fields, their titles, brief definitions, and data types that make up the Chronic Kidney Disease dataset. There are a total of 25 characteristics and 400 samples in the dataset. Class is the area of focus for disease forecasting, and it has been implemented. Eleven are numerical and fourteen are category, making up the remaining 24 attributes.

3.1.1 Clinical Data Preprocessing:

Pre-processing steps include looking for and eliminating outliers, validating and normalizing noisy data, and checking for missing values. The patient evaluation reveals several missing or inaccurate estimates. The planned research endeavor includes the following pre-processing measures to account for this.

3.1.1.1 Handling of Missing Values

The simplest approach to dealing with missing values is to ignore the records, although this is impractical for smaller datasets. The data set is checked for missing attribute values as part of the data production process. By using the statistical technique of mean imputation, we may assess the value of missing data in numerical systems. When filling up missing values for minor characteristics, the mode technique is employed.

3.1.1.2 Categorical Encoding of Data

Since most deep learning methods only take numbers as input, the category values must be converted into numbers. Yes and no are only two examples of the types of things whose attributes are represented by the binary integers 0 and 1.

3.1.1.3 Transformation of data

Data transformation refers to the act of changing numbers on a small scale such that one variable does not dominate the others. Otherwise, regardless of the unit of measurement, the learning methods will consider larger values to be more advanced and smaller ones to be less advanced. For further processing, the data modifications adjust the values in the dataset. This study employs a data normalization technique to improve the precision of deep learning methods. Transformed information has a standard deviation of 1, a mean of 0, and an intermediate value of -1.

The standardization can be stated as,

$$s = \frac{(v - \bar{v})}{\sigma} \quad (1)$$

The above equation denotes the standardized score, the observed value is represented as, the mean is denoted as, and signifies the standard deviation.

3.1.1.4 Outlier detection

In statistics, outliers are single observations that can't be explained by the rest of the data. The estimate of experimental variability or signal inaccuracy might generate an outlier. Outliers may skew and mislead a deep learning algorithm's training data. Longer training times, less reliable output, and reduced model accuracy are all consequences of outlier inclusion. This study employs an IQR-based technique to filter out extreme cases before feeding the data into the learning algorithm.

When a dataset is divided into quartiles, the interquartile range (IQR) may be used as a measure of variability. The numbers that separate each section are known as the first, second, and third quartiles, and the formula for determining the interquartile range (IQR) is as follows.

$$IQR = V_3 - V_1 \quad (2)$$

where V_3 denotes the middle value in the first half of the ordered data set whereas V_1 denotes the second half. V_1 denotes the median value in the dataset.

3.2 Imaging Data

Computed tomography (CT) scans were used to acquire the imaging data. The collection includes several different kinds of pictures. The pictures were taken using standard techniques and saved in DICOM format for uniformity and ease of analysis.

3.2.1 Data Pre-Processing

Extensive preprocessing processes were conducted on the clinical and imaging data to assure quality and compatibility.

First, the input CT images are class labeled and then pre-processed to remove noise. To restore integrity to characteristics of an image that have been degraded by noise, a pre-processing technique is presented for use in noise reduction. In this case, adaptive filtering is used for local denoising of an image's noise content. One definition of a corrupted picture is one in which by $\hat{I}(x,y)$, the noise difference across the whole image is shown by σ_y^2 , the local mean is provided by μ_L . Approximately apixel window and local variance window is provided by σ_y^2 .

4. Feature Extraction and Classification:

4.1 Clinical data

4.1.1 Estimation of Glomerular filtration rate

The GFR is very important in many contexts, including as public health, medical treatment, and scientific inquiry. Clinical labs play a crucial part in determining GFR and diagnosing chronic renal disease. Serum creatinine measurement and GFR estimation are advised as the first steps in the GFR assessment. Here, the filtration rate is determined using age and serum creatinine (SC) level, and the GFR is used to categorize the five phases of CK illness.

4.1.2 Classification:

Graph neural network with deep Q learning technique:

GNNs are the framework together the node dependency in graphs through the passing of messages between the nodes. The GNN performs on the graph to describe the data from its neighborhood with random stages. This creates GNN as an appropriate tool to utilize for wireless networks that hold compound features that cannot be taken in a closed form. In the proposed research work, the GNN-based approach in accordance with the Q-function is learned from the cell and entity placement instances through a deep Q-learning approach. The major merit of the Q-function is to establish GNN scalable over various sizes that can gather limited network features with different numbers of cells and entities. To generate the optimal selection, the right Q function has to be learned. When the Q function is gathered through GNN, this renders to learn the GNN parameters which is done by sequential

accumulation of new cell entity connections over partly connected graph, with the relationship of cells and entities between the nodes.

4.2 Image Data

4.2.1 Feature extraction using fusion model

Here, pre-processed CT scans have been fed into VGGNet-16 and Inception v4 feature extractors for the purposes of COVID-19 prediction and classification, respectively. The classifier receives a combined set of extracted characteristics from these models.

The convolutional neural network (CNN) model uses a series of convolution layers to identify preexisting picture patterns. CNN's strength is in its ability to train an extremely deep network with just a small number of parameters. It makes the training process easier and takes less time. The CNN also has many levels, such as the convolutional layer, the activation layer, the pooling layer, the Fully Connected layer, and the SoftMax layer.

VGGNet-16 model: In 2014, the Oxford Visual Geometry Group introduced VGG-16, a popular CNN technique with 16 layers that has shown standard results across a variety of image processing applications (Xu et al., 2019). When increasing a system's depth, VGG16 swaps out large convolution filters for small ones. The improvement in classification accuracy is mostly attributable to CNNs with very small filters. Front-layers of the pre-trained CNN methodology are an applicable low-level universal feature that is suited for normal image processing applications, and the VGG-16 CNN method used in this work was pre-trained on the ImageNet dataset. New version of the Inception algorithm. Inceptions are utilised throughout several training stages to break up repetitive blocks into smaller networks more suited to displaying a whole model in memory. Therefore, Inception modules are simply tweaked, representing the potential of altering the number of filters from exclusive layers, without affecting the preeminence of the trained network. The training time may be reduced by fine-tuning the layer size to strike a balance between the various sub-networks. The newest Inception models were developed using TensorFlow, and unlike previous models, they do not include any redundant segmentation. This may be due to the fact that activating tensors, which play a crucial role in calculating gradient and approximating bounded values, is a function of contemporary memory optimisation for Back Propagation (BP). Furthermore, inception-v4 is planned to remove duplicated work across all grid sizes for Inception blocks (Shankar et al., 2020).

Residual inception blocks: The filter-expansion layer, which is used to increase the filter bank's dimensionality, applied inception blocks in this model before calculating the input depth. Important for making up for the reduction in dimension necessitated by the Inception block. Inception-v4 is mild because to the inclusion of several levels, and it is only one of many Inception variants. The supplementary shift between residual and non-residual forms of change. For typical layers, we make use of something called Batch Normalisation (BN). Since the BN model in TensorFlow uses more memory, it's important to restrict its use to certain situations and reduce the total number of layers.

Scaling of the residuals: Here, if the number of filters is more than 1000, the network is terminated during the first stage of training, which is the destination layer before the pooling layer activates to construct zeros from different iterations, since residual techniques reveal its instability. As a result, restricting training methods won't be enough to get rid of it. The learning method is also confirmed to be effective when the constrained measurements added before the activation layer are used. Accumulated layer activations are typically scaled using factors between 0.1 and 0.3.

4.2.2 Classification:

The suggested ensemble deep learning model that makes use of the Internet of Things is illustrated below in stepwise. In the beginning, medical IoT devices get the necessary scan of a patient at a nearby medical facility. After a scan is completed, it is sent to the IoT framework's storage layer over a communication medium. The resulting scan is then processed using the ensemble deep learning model. Once the results have been received, they are saved in the database. Medical specialists, physicians, and patients are just some of the IoT users who may have access to data stored in the cloud. Figure 3 depicts the suggested ensemble deep learning model. It makes it quite apparent that we'll be training the pre-trained models independently at first. The final ensembled framework for automated screening of CKD suspicious instances is obtained by a majority vote. The remainder of this article will focus on the sequential ensemble model.

- 1 Initially, the abdomen CT dataset is obtained
- 2 Divide the dataset into training and testing fractions, i.e., 65% and 35%, respectively.
- 3 Mathematically, each set is defined as

$$[C_{tr}C_{ts}] = T_f(D_s). \quad (1)$$

Here, C_{tr} represents the training set of CT scans, C_{ts} shows the testing set, T_f is the 10-fold cross-validation, and D_s is the collected four-class CT scan dataset.

- 4 The deep learning models, i.e., ResNet152V2, DenseNet201, and IRNV2 are applied on the testing dataset (C_{ts}) as

$$\begin{aligned} R_s &= T_L(R, S), \\ D_s &= T_L(D, S), \\ I_s &= T_L(I, S). \end{aligned} \quad (2)$$

Here, R_s , D_s , and I_s show the softmax functions of ResNet152V2, DenseNet201, and IRNV2, respectively, T_L represents the deep transfer learning model, R , D , and I demonstrate the ResNet152V2, DenseNet201, and IRNV2, respectively, and S shows the softmax function.

- 5 The trained individual deep transfer learning models can be defined as

$$\begin{aligned} R_s &= M_B(R_s, C_{tr}), \\ D_s &= M_B(D_s, C_{tr}), \\ I_s &= M_B(I_s, C_{tr}) \end{aligned} \quad (3)$$

- 6 Here, M_B defines the model-building process. Finally, ensembling is achieved by using majority voting as

$$E_C = E_M(R_s, D_s, I_s) \quad (4)$$

Here, E_C is the trained ensemble diagnostic model. E_M defines the majority voting ensemble model.

4.3 Fusion Methodology

In order to improve the precision and reliability of kidney disease detection, the fusion technique used in this study combines clinical and imaging data in a complementary fashion. The fusion methods and algorithms explored in this study are outlined below.

4.3.1 Decision-Level Fusion

Combining the predictions or choices generated by separate models trained on clinical and imaging data is the subject of decision-level fusion. Using methods of ensemble learning including majority voting and weighted average, this study fuses individual decisions. The final determination is based on the combined predictions of many classification models that were trained independently on clinical data and imaging data. Combining these verdicts strengthens the renal disease detection method by compensating for the shortcomings of individual models.

When combining the results of many models or modalities into a single conclusion, "max voting" is a strategy that favors the model or modality with the most votes or the greatest confidence score. Ensemble learning and the merger of several classifiers use this straightforward and instinctive method. To get a consensus, max voting takes into account predictions from several models or modalities and then chooses the forecast with the most votes. Multiple classes may be tied, in which case many methods exist for breaking the tie, such as choosing the class with the greatest average confidence or using a priority order for the classes. Max voting's merits are its simplicity, robustness, and readability. Equal weighting, a lack of uncertainty estimates, and how ties are handled are just a few of the concerns and constraints that must be taken into account while using max voting.

In cases when the models or modalities have comparable performance and no substantial variances in their expertise, max voting may be an effective decision-level fusion strategy. However, before using max voting or any other fusion approach, it is crucial to analyse the unique features of the issue, the performance of individual models, and the desired qualities of the integrated system.

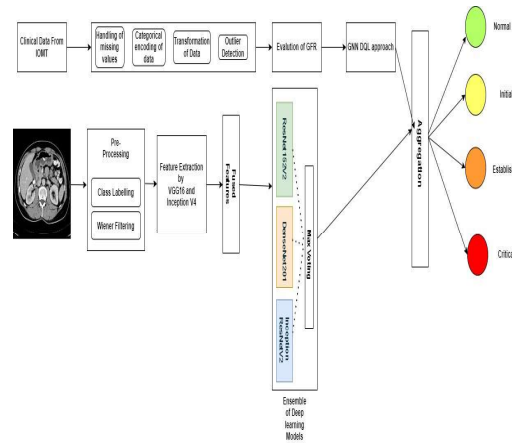


Figure 3. MultiModal Deep Fusion binary classification model for CKD

5. Results

The CPU requirement for augmentation and classification jobs is Intel i7-9700F with 16GB of RAM, and the models have been trained on NVIDIA GeForce RTX-2060 SUPPER with 8GB of memory. Python 3.8.5 and the Tensorflow 2.5 package for DL models are used to implement the research.

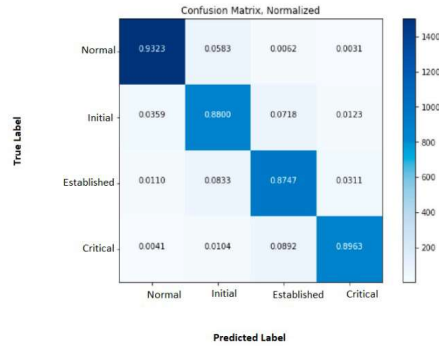


Figure 4. Confusion matrix

Table 1. Metrics of performance

Models	Performance outcomes (%)			
	Accuracy	F1 score	Recall	Precision
GNN-DQL (Clinical Data)	99.93	99.86	99.86	99.86
Ensemble Model (Image Data)	98.05	98.24	98.05	98.43

The figure 4 is the confusion matrix of the multi-classification and the performance metrics are calculated with the displayed data in the confusion matrix. The performance metrics are calculated separately for clinical data model and image data model as shown in the table 1, and then the decisions are interleaved to attain the complete prediction.

Table 2. Benchmarking with Human expert

Evaluation Factors	Human Expert	Proposed Model
Accuracy	0.960	0.960
Recall	1.000	0.917
Precision	0.923	1.000
F-1 Score	0.960	0.957

6. Conclusions

Table 2 shows that when clinical data is combined with the CTD, the system outperforms a human expert in making diagnoses. The suggested system will not be able to diagnose CKD on its own, but it will help physicians in countries where it is difficult to get medical professionals. Despite the fact that the suggested approach has yielded very substantial outcomes. There are flaws in the research as well. The validation was carried out on a sample size of 250 patients since only one expert was available and a significant number of patient data were not available.

References

- [1] Tai, Y., Gao, B., Li, Q., Yu, Z., Zhu, C. and Chang, V., 2021. "Trustworthy and intelligent covid-19 diagnostic iomt through xr and deep-learning-based clinic data access". *IEEE Internet of Things Journal*, 8(21), pp.15965-15976.
- [2] Elbasi, E. and Zreikat, A.I., 2021, May. "Efficient early prediction and diagnosis of diseases using machine learning algorithms for IoMT data". In *2021 IEEE World AI IoT Congress (AIoT)* (pp. 0155-0159).IEEE.
- [3] Akter, S., Habib, A., Islam, M.A., Hossen, M.S., Fahim, W.A., Sarkar, P.R. and Ahmed, M., 2021. "Comprehensive performance assessment of deep learning models in early prediction and risk identification of chronic kidney disease". *IEEE Access*, 9, pp.165184-165206.
- [4] Yu, C.S., Lin, Y.J., Lin, C.H., Lin, S.Y., Wu, J.L. and Chang, S.S., 2020. "Development of an online health care assessment for preventive medicine: a machine learning approach". *Journal of medical Internet research*, 22(6), p.e18585.
- [5] Mitchell, H.B.: *Data Fusion: Concepts and Ideas*. Springer, Heidelberg (2014).doi:10.1007/978-3-642-27222-6
- [6] Lahat, D., Adali, T., Jutten, C.: *Multimodal data fusion: an overview of methods, challenges, and prospects*. *Proc. IEEE* 103(9), 1449–1477 (2015). doi:10.1109/JPROC.2015.2460697Fusion of Clinical Data: A Case Study 301
- [7] Castebedo, F.: A review of data fusion techniques. *Sci. World J.* (2013). doi:10.1155/2013/704504
- [8] Rohlfing, T., Pfefferbaum, A., Sullivan, E.V., Maurer, C.R.: Information fusion inbiomedical image analysis: combination of data vs. combination of interpretations.In: Christensen, G.E., Sonka, M. (eds.) *IPMI 2005*. LNCS, vol. 3565, pp. 150–161.Springer, Heidelberg (2005). doi:10.1007/11505730_13
- [9] Ponti Jr., M.P: Combining classifiers: from the creation of ensembles to the decisionfusion. In: *24th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials(2011)*
- [10] Lee, G., Madabhushi, A.: A knowledge representation framework for integration,classification of multi-scale imaging and non-imaging data: preliminary results inpredicting prostate cancer recurrence by fusing mass spectrometry and histology.In: *International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE(2009).Irvin, J. et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *ArXiv190107031 Cs Eess* (2019).
- [11] Rajpurkar, P. et al. CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *ArXiv171105225 Cs Stat* (2017).
- [12] Bien, N. et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLOS Med.* 15, e1002699 (2018).
- [13] Rajpurkar, P. et al. AppendiXNet: Deep learning for diagnosis of appendicitis from A small dataset of CT exams using video pretraining. *Sci. Rep.* 10, 3958 (2020).
- [14] Yang, X. et al. A two-stage convolutional neural network for pulmonary embolism detection from CTPA images. *IEEE Access* 7, 84849–84857 (2019).
- [15] Huang, S.-C. et al. PENet—A scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. *NPJ Digit. Med.* 3(1), 1–9 (2020).
- [16] Tajbakhsh, N., Gotway, M. B. & Liang, J. Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*, Vol. 9350 (eds Navab, N. et al.) 62–69 (Springer, Berlin, 2015).
- [17] Leslie, A., Jones, A. J. & G& Goddard, P. R. Te influence of clinical information on the reporting of CT by radiologists. *Br. J. Radiol.*4, 1052–1055 (2000).