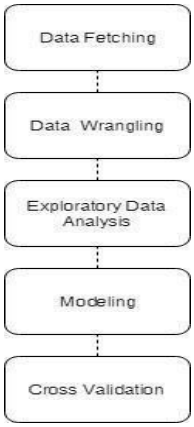
IMPLEMENTING DATA ANALYSIS TO ASSOCIATE KIDNEY AND HEART DISEASES USING MACHINE LEARNING ALGORITHMS

'P Saranya', 'Manan Cholera', ' Siddharth Nepak'

Abstract- **Big data presents large quantities of data in various formats and in order to extract maximum useful information from the dataset, a number of data integration techniques like data fusion can be used. Data is normalized before the actual analysis begin. By doing so, we attain a linear distribution of components from the dataset when we perform PCA (Principal Component Analysis). It is an algorithm used to reduce the dimensionality of the dataset and create PCA components which would orthogonally portray a relationship between all the other components of the dataset so that when there is a new patient entry in the dataset, it can easily categorize the outcome of the new entry based on the PCA component. In order to have better accuracy of our analysis, a model is trained to predict the outcome of our analysis. After comparing the accuracies of different algorithms, a model consisting of common components with similar characteristics between the datasets is trained and existing machine learning algorithms are applied to predict if a patient having chronic kidney disease symptoms suffers from heart disease as well by categorizing the outcome into four different stages. After applying multiple algorithms, the algorithm with the most reduced error-rate, and the best accuracy to perform the data analysis to associate chronic kidney disease and heart disease are selected**.

I. INTRODUCTION

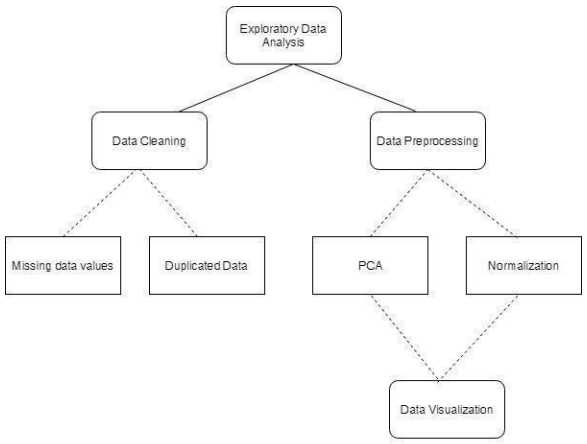


Before beginning with the data analysis process, it is really important to clean the data that we will be using for our process. There are numerous ways to perform to clean the data and transform it with the intent of making it more appropriate to work with and make more accurate.

Data Wrangling maps data in one format to another format in order to make it more valuable for a variety of purposes such as analytics, training-testing process, visualization of data etc. These data transformations are usually applied to different entities of the dataset to create desired wrangling output which could be used for further analysis of our dataset. In order to clean the data, various techniques such as filling the missing values, handling incorrect data types and duplicated data etc can be used. After cleaning the dataset, data analysis by applying different functions using various python libraries and the result of our analysis is displayed using different data visualizations methods.

A majority of datasets consist of a lot of components and rescaling the data becomes a priority. One single entry in the dataset will have projections from all the components in the dataset and predicting if a patient has the disease or not becomes difficult. In order to improve the outcome of our analysis,

we normalize the data before beginning with data analysis. StandardScaler preprocessing module is used to normalized the features, which sets the 'mean=0' and 'variance=1' for all the features. If you normalize the features, it would provide a fair comparison between the explained variance in the dataset. The overall performance of the PCA is dominated by high variance features. Therefore, features should be normalized.



ARCHITECTURE DIAGRAM

The PCA components can only be numeric. So the common way is to encode the categorical values by using one-hot encoding or any other encoding method to transform the values into 0 and 1. PCA looks for the correlation between the features and reduces the dimensionality. The same process is repeated for both the datasets and the outcome of our analysis is portrayed using visualizations.

PCA is an unsupervised machine learning method and hence we use this method only for exploratory data analysis purpose in order to gather insights about the dataset. Next, a supervised machine learning algorithm is used to train a model for predictive analysis. As usual, the dataset is processed first and a neural network is formed where the components with common characteristics between the datasets are given as input. DNN is trained, tested and result is displayed with accuracy as the metric. Compare the result from the DNN to another ML algorithm.

II. BACKGROUND

Challenges are faced in nearly every data-driven projects and some of them are Data quality, Merging and Linking, Big data, Dirty data, uncertainty, error tolerance etc but it is still a challenging process. A number of solutions exist for single problems, none of them really solves our problem. Different concepts and tools used for data wrangling require additional insights.

PCA is a reduction technique which combines the input variables in a specific way, then drops the "least important" variables and retains the most valuable parts of all the variables. Each of these "new" variables are independent of each other. This is an added benefit because the assumptions of a linear model require the variables to be independent of one another.

III. RELATED WORKS

1. This is a work in which Bayesian Filtering technique is used for the Location tracking system. The computational load of the algorithm is reduced with the aid of different observation for location tracking. The advantages are it's less computational complexity and high accuracy whereas the disadvantages includes that the incoming information may not be always highly reliable.

2. This is a work uses Kalman Filter in order to work on rail track monitoring system. It enables data-driven rail-infrastructure monitoring from multiple trains. The advantages includes that it handles asynchronous data and increases overall reliability of inspection. The disadvantages for the same are high deduction cost and that it works with high cost censors.

3. In this work Kalman Filter data fusion techniques is used for vehicle localization. It identifies and recovers hardware and software faults and makes a fault tolerant architecture. The advantages are that it focuses on both hardware and software faults and it has no false positives. The limitations include it's time consuming nature and less accuracy (62.32% for hardware faults and 90% for software faults).

4. In this piece of work the technique used is double sample data fusion. The purpose of this work is fault diagnosis in petrochemical rotating machinery equipment. The advantage is that it has 9.45 % of accuracy improvement when compared to KNN combination method and the disadvantage is that it's reliability calculation method is inadequate.

5. This is a piece of work is done using the feature based data fusion technique. The purpose is for dwelling detection in Refugee camps. It has a high accuracy, no false positives, takes less time and works with a complex structure. Automatic detection can be for the improved in this piece of work.

6. This uses feature level fusion and SVM method. It is used for human activity recognition in Health Care Centers. It has a high increased accuracy of 98.32% but it only works for single structure and not for multiple structures at once.

Sl No	Paper	Author	Technique used	Advantages	Disadvantages	Application	Purpose
1	A reduced complexiaty data fusion algorithm using belief propagation for location tracking in heterogeneous observations	Yih-Shyh Chiou and Fuan Tsai	Bayesian Filtering	Less computational complexity, High accuracy	Incoming information may not always be highly reliable	Location tracking system	To reduce the computational load of the traditional data-fusion algorithm with heterogeneous observations for location tracking.
2	A data fusion approach for track monitoring from multiple in-service trains	George Lederman, SihengChen, JamesH.Garrett, Jelena Kovacevic, Hae Young Noh, Jacobo Bielak	Kalman Filter	Handles Asynchronous data, Increases overall reliability of inspection.	High detection cost, Works with high cost sensors.	Rail track monitoring system	Enabling data-driven rail-infrastructure monitoring from multiple in-service trains.
3	A fault tolerant architecture for data fusion: A real application of kalman filters for mobile robot localization	Kaci Bader, Benjamin Lussier, Walter Schon	Kalman filter Data Fusion	Focusing both hardware and software faults, No false positives	Time consuming, Accuracy is less(62.32% har dware faults, 90% Software faults)	Vehicle Localization	Identification and recovery of hardware and software faults (fault tolerant architecture)
4	Double sample data fusion method based on combination rules	Jianbin Xiong	Double sample data fusion method	9.45% of accuracy improvement when compare to KNN combination method	Reliability calculation method is inadequate	Petrochemical rotating machinery equipment	Fault diagnosis
5	Object-Based Analysis and Fusion of Optical and SAR Satellite Data for Dwelling Detection in Refugee Camps	Kristin Sprohnle, Eva-Maria Fuchs and Patrick Aravena Pelizari	Feature based data fusion	High accuracy (nofalse positives), Less time, Works with complex structure	Automatic detection need to be improved	Refugee camps	Dwelling detection
6	A Data Fusion-Based Hybrid Sensory System for Older People's Daily Activity and Daily Routine Recognition	Yan Wang,Shnang Cang and Hongnian Yu	Feature level fusion and SVM method	Increased accuracy of 98.32%	Works for only single structure(not for multiple)	Health care	Human activity recognition

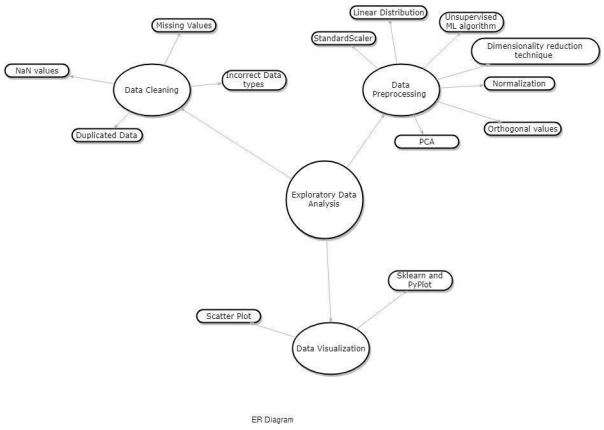
Literature Survey

IV. METHODOLOGY

In spite of the advances in recent trends in technologies, data analysts still spends a huge amount of time identifying the data quality issues and converting the data into a much more usable form. Even though data cleaning and data integration are the longstanding issues, relatively little research has shown how interactive data visualization can advance the state of our data analysis.

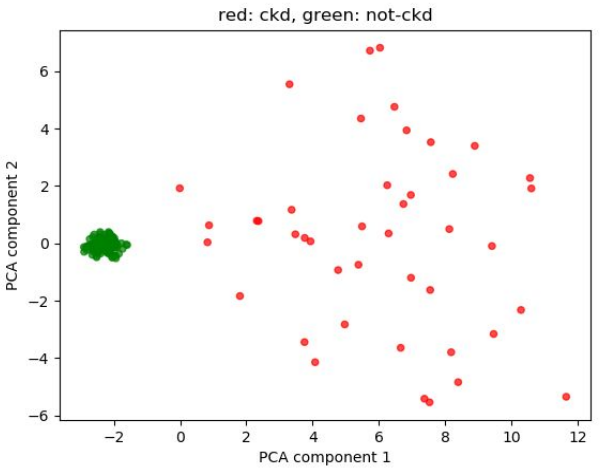
If we wish to fit a regression model using these principal components, these assumptions will be satisfied. PCA considers the total variance in the data and transforms

the original set of variables into a smaller set of linear combinations to perform factor analysis.

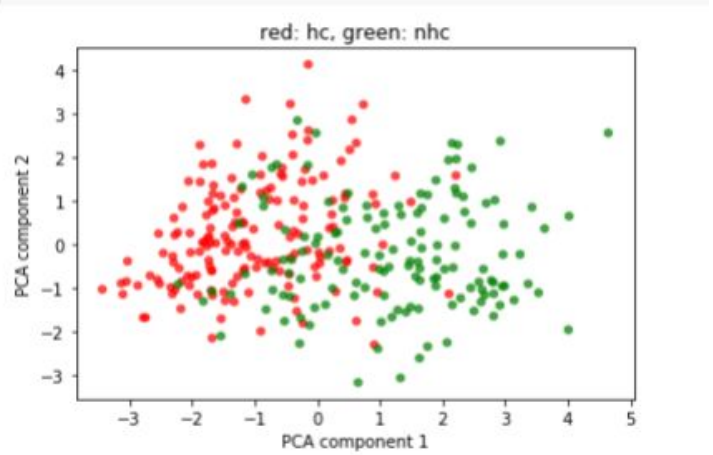


ER Diagram

PCA is generally used when the most important thing is to determine the minimum number of factors that will be responsible for variance in the dataset. The PCA components should be numeric. So the common way is to encode the categorical values by using one-hot encoding or any other encoding method to transform the values into 0 and 1. PCA is an unsupervised machine learning method and hence we use this method only for exploratory data analysis purpose in order to gather insights about the dataset. The main purpose of using PCA is for identification of different patterns linking the attributes in the dataset. For the PCA model for chronic kidney disease and heart disease, data is read and then cleaned by dropping the NaN values and color-labeling the outcome of the target values which represent the outcome of the analysis. In order to perform PCA, all the categorical components in the dataset are converted to numeric components by using encoding techniques such as one-hot encoding. For every column with categorical values, two copies are created with entries 0 and 1 and the original column with the categorical value is dropped. StandardScaler is used to rescale the components. The number of components are defined and the data frame is fitted in the model. The outcome of our analysis is plotted on a graph to show the results.

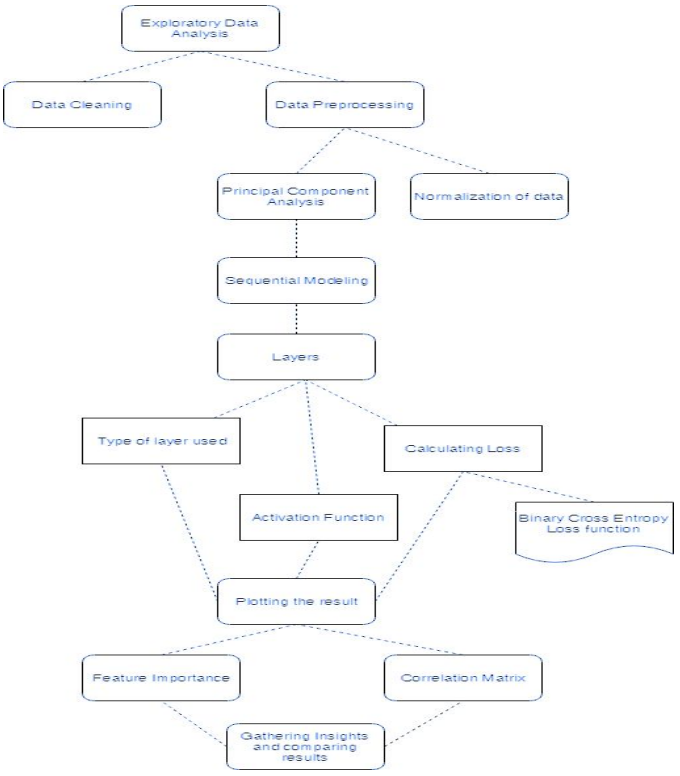


PCA for Chronic Kidney Disease



PCA for Heart Disease Prediction

V. MODEL ARCHITECTURE



System Architecture Diagram

Neural networks are well known for classification problems. In order to use a DNN for predictive analysis, steps involved are :

1. Process the dataset.
2. Make the deep neural network.
3. Train the DNN.
4. Test the DNN.
5. Compare the result from the DNN to another ML algorithm.

Preprocessing of data is minimal. Features with missing values are dropped and the categorical features are encoded. The first step in making the deep neural network is defining a model for training our data.

The models which we will be using here are the keras models. They were released in 2015 with an aim to enable a fast experimentation Keras uses 'model' data structure to organize the layers. There are many types of models available and the simplest type of keras model is the sequential model. It consists of layers piled up one on another. Models use the Keras API for complex architectures. The sequential one allows the users to create layer-by-layer models for almost all of the problems. The only drawback of keras is that it does not allow the creation of models that share layers. The Keras API is actually used as a substitute method for creating such models that offer significantly greater capability to adapt to the existing complexities, including making progressively complex models. The error rate of our analysis is decreased and the detection of the error is much more easier.

Various steps involved in Sequential Keras are:

1. LoadData.
2. Define Model.
3. Compile Model.
4. Fit Model.
5. Evaluate Model.
6. Tie It All Together.

A dense layer is only a regular layer of neurons in a neural system. Every neuron receives instructions and input from every one of the neurons in the previously present layer, in this manner densely associated.

The layer has the following attributes :

- a weight matrix 'W',
- a bias vector 'b',
- the activation of former layers 'a'.
- Output = activation(dot(input, kernel) + bias).
- **activation** is the element-wise activation function passed as the activation argument,
- the **kernel** is a weights matrix created by the layer,
- **bias** is a bias vector created by the layer.

Code Snippet:

```
- from keras. Models import Sequential
- from keras. Layers import Dense
- model= Sequential ()
- model.add(Dense(2, input dim=1))
- model.add(Dense(1, activation_function))
```

The main steps involved in making a DNN are :

1. Define a sequential model.
2. Add some dense layers.
3. Use an activation function for the hidden layers.
4. Use a normal initializer as the kernel_initializer.
5. Use a binary_crossentropy as the loss function.
6. Define the output layer with only one node.
7. Use the sigmoid function for the output layer.

We use the activation function with the neural network to determine the output of the neural network.

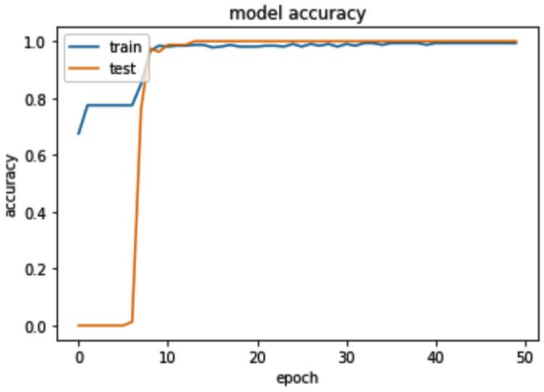
The Non-Linear activation functions are the most used activation functions as these functions make it simple for our model to use the data to its full potential. The different types of activation function available for use are :

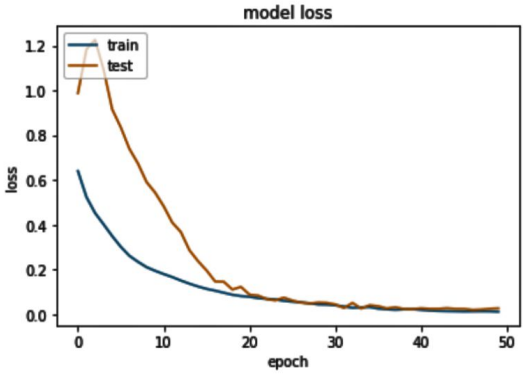
1. 'Sigmoid' function
2. 'tan-h' function
3. 'ReLU' function
4. 'Leaky ReLU' function

.

- **Sigmoid** - The primary reason for using this function is because of its range (0-1). It is usually used for models where the probability is predicted for the output.
- **tan-h** – tan-h is like a logistic sigmoid function but better in performance. The range of a typical tan-h function is from -1 to 1.
- **ReLU** - It is the most used activation function. But the only problem is that all the values less than zero will become zero which affects the model's ability to train the data properly. That means that the actual graph would be different from the correct output.
- **Leaky ReLU** - It is a function used to solve the ReLU function problem. The leak actually helps in increasing the range of the ReLU function.

The loss function is one of the two parameters needed to fit, train and compile a model. **binary_crossentropy** is the loss function used here. The model accuracy and model loss are calculated and visualized on a graph.





The accuracies of our outcome are compared with accuracies using different ML algorithms. Next step is to identify the similar characteristic components from both the datasets and compare them with their threshold values. The outcome of our analysis will be able to tell us if a person having a chronic kidney disease may or may not have heart disease. The ML algorithm with the best accuracy is chosen for analysis.

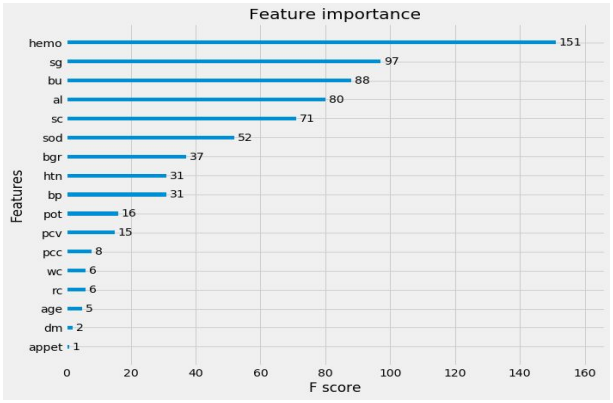
For the chronic kidney disease model, an accuracy of 99.50% was reached based on the outcome of 400 epochs. For the heart disease predictive model, logistic regression and random forest classifier is used wherein logistic regression had an 86.89% accuracy and random forest classifier had an accuracy of 88.52%.

Logistic Regression is a Statistical Learning technique categorized in Supervised Machine Learning methods dedicated to Classification tasks. It is recognized as one of the most used machine learning algorithms for binary classification because this algorithm is very simple to use and also performs well on a variety of problems. RFC is an ensemble algorithm which are those that combine multiple algorithms of different or same kind for classifying various objects. These trees are generated from a randomly selected subset of the training dataset. All our models work fine but best of them is the RFC with an accuracy of 88.52%.

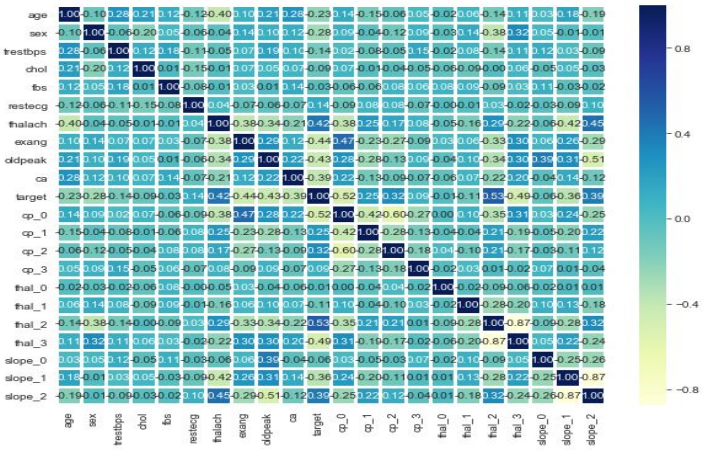
For chronic kidney disease, we will use the feature importance algorithm which uses the Extreme Gradient Boosting Classifier. A feature is important as the prediction error of the model increases after the permutation of the values of different features of the dataset. This breaks the relationship between the true outcome and the feature. The importance of a feature is measured by calculating the increase in the model's prediction error. A feature can either be 'important' or 'unimportant'. A feature can be called "important" if shuffling its values increases the model error and a feature is called "unimportant" if shuffling its values leaves the model error unchanged.

XGBoost is an applied machine learning algorithm used for structured or tabular data. Gradient boosted decision trees are implemented using this algorithm It is specifically used for its speed and performance. The model supports the features of the R implementations and the scikit-learn, with new features like regularization of data. The implementation of this algorithm was designed for efficiency of memory resources, compute time etc. XGBoost is really fast when compared to other

implementations of gradient boosting. F-Scores are a measure of how informative each feature is for the dataset.



Feature Importance Plot using xgb Classifier for Chronic Kidney Disease model



Correlation Matrix for Heart Disease Prediction

After comparing the models for heart disease prediction, we will then use a correlation matrix to determine the correlation between various features with the target value in the dataset. The association between random variables is a structured measure or a 'correlation'. There are various methods to calculate the correlation coefficient, each measuring different types of strength of association. The variables in our dataset may be complexly related with one another and thus it becomes important to quantify the degree to which these variables are interdependent.

So from this correlation matrix, we can observe which input features contribute most during the training process of various classifiers. Blood disorder , heart rate slope, maximum heart rate and chest pain types can be used for prediction more accurately than other features.

VI. CONCLUSION

These datasets are old and small by today's standards. However, a simple model is created and then various machine learning explainability tools and techniques are used. At the start, it was hypothesised, using domain knowledge that factors such as cholesterol, blood pressure and age would be major factors in the model for heart disease prediction. This dataset didn't show that. Instead, the number of major factors and aspects of ECG results dominated. For Chronic Kidney disease prediction, Feature Importance algorithm is used which shows the increase in the percentage of model error when the information about the feature is destroyed. The results were different from what the theoretical knowledge suggested the results would be. This sort of approach will become increasingly important as machine learning has a greater and greater role in health care.

REFERENCES

1. Florian Endel, Harald Piringer : Data Wrangling: making data useful again.

2. Dr. Rama Kishore, Taranjit Kaur : Backpropagation Algorithm: An Artificial Neural Network Approach for Pattern Recognition .

3. Cristinel Constantin : Principal Component Analysis - A powerful tool in computing marketing information.

4. Y. Chiou and F. Tsai, "A Reduced-Complexity Data-Fusion Algorithm Using Belief Propagation for Location Tracking in Heterogeneous Observations," in IEEE Transactions on Cybernetics, vol. 44, no. 6, pp. 922-935, June 2014.

5. George Lederman, Siheng Chen, James H. Garrett, Jelena Kovacevic, Hae Young Noh, Jacobo Bielak, “A data fusion approach for track monitoring from multiple in-service trains,”Mechanical Systems and Signal Processing,Volume 95, 2017.

6. B. Kaci, L. Benjamin and S. Walter, "A Fault Tolerant Architecture for Data Fusion Targeting Hardware and Software Faults," 2014 IEEE 20th Pacific Rim International Symposium on Dependable Computing, Singapore, 2014

7. J. Xiong, Q. Zhang, Z. Peng, G. Sun and Y. Cai, "Double Sample Data Fusion Method Based on Combination Rules," in IEEE Access, vol. 4, pp. 7487-7499, 2016.

8. K. Spröhnle, E. Fuchs and P. Aravena Pelizari, "Object-Based Analysis and Fusion of Optical and SAR Satellite Data for Dwelling Detection in Refugee Camps," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 5, pp. 1780-1791, May 2017.

9. Y. Wang, S. Cang and H. Yu, "A Data Fusion-Based Hybrid Sensory System for Older People’s Daily Activity and Daily Routine Recognition," in IEEE Sensors Journal, vol. 18, no. 16, pp. 6874-6888, 15 Aug.15, 2018.