A Machine Learning Technique to Detect Counterfeit Medicine Based on X-Ray Fluorescence Analyser

Muna Alsallal
STI
ATU Technical University
Baghdad, Iraq
ins.mun@atu.edu.iq

Mhd Saeed Sharif
School of Architecture, Computing and Engineering,
College of Arts, Technology and Innovation, UEL.
University Way, Dockland Campus
London, E16 2RD, UK.

Baydaa Al-Ghzawi
STI
ATU Technical University
Baghdad, Iraq
baydaafawz@gmail.com

Sabah Mohammed Mlkat al Mutoki
STI
ATU Technical University
Baghdad, Iraq
asabah_sh200@yahoo.com

***Abstract*— Since so many sub-standard and fake medicines are being openly sold, the counterfeit medicines have become widespread. The forgers succeeded in imitating the genuine medicines and make them look like genuine ones. This paper has proposed an approach that based on analysing the Tenormin® 50mg medicine by using non-destructive X-Ray Fluorescence Technique. This technique has been proposed over other heavy chemical analyzing methods to detect counterfeit Tenormin® due to its speed and reliability. There are 10 samples of Tenormin tablets from different manufactures were tested. All samples contained the active element Atenolol 50 mg and other inactive elements. Moreover two supervised machine learning techniques; RBF Support Vector Machine (RBF-SVM) and K-Nearest Neighbor (KNN) are employed. These two supervised machine learning algorithms were proposed as a step to design an automated approach in order to determine fake from genuine Tenormin® without a need for trained chemists. The results revealed that X-Ray Fluorescence Technique has discriminated three elemental composition samples which differ from other 7 samples. The results also revealed the SVM proposed approach outperforms the KNN based approach with an overall accuracy of 93%.**

Keywords—counterfeit-medicines, XRF-Minipal2, SVM, KNN, elemental-composition

I. INTRODUCTION

Nowadays, counterfeit medicines are being a problem in less developed countries, this can be attributed to the ineffective deterrence laws, the absence of drug control roles and the lack of community awareness. The main sources for such counterfeit medicines are India and China as stated by [1]. In developed countries, the new life style has encouraged people to buy medicines through the internet and websites [1]. The world health organisation has defined the counterfeit medicine as "counterfeit products may include products with the correct ingredients or with the wrong ingredients, without active ingredients, with insufficient active ingredient or with fake packaging." [2]. According to WHO (The World Health Organization) around thirty five percent of medicinal market in middle east region, in particular Arabic countries assumed as illegal [3]. That indicates that this region is vulnerable to oversupplied with counterfeit medicines that contain no active elements. Such counterfeit medicines can include harmful materials such as rat poison and cartridge ink [3]. The harmful medicines lead to severe side effects like brain damage and heart problems, in addition to other side effect that highlighted by healthcare researchers [3].

Towards building a robust system to combat the spread out of counterfeit medicines, the UAE announced during the 2nd Emirates international symposium on fighting fake medicines that it is planning to use technology to detect such products. The technology revolves around using a smart device that can detect the fake medicine within seven seconds [3]. Such devices can be able to do the mission if they rely on non-destructive checks to characterize the elemental composition such as X-ray fluorescence technique (XRF) [4].

The technique of XRF (X-ray fluorescence) is a non-destructive analytical technique that can performs procedure to define the materials' elemental composition [5]. This device is known as an analyser that can measure the luminous (fluorescent) X-ray that produced from a test sample when it is manipulated by a primary X-ray source in order to determine the test sample chemistry [6]. XRF-Minipal2 also identified as a device that can provide semi-quantitative or qualitative facts about the main elements [6, 7]. X-ray fluorescence (XRF-Minipal2) has been reported by [11, 12] as an appropriate method for classification of the metal presence.

This paper proposed an integrated technique that combine between the X-ray fluorescence technique (XRF-Minipal2) as an elemental composition and machine learning (ML) algorithm to replace human for classification tasks.

The proposed technique works on mechanism of obtaining the analysed data of organic and counterfeit elements of Tenormin drugs to be used as input for machine learning algorithm. This technique can be considered as a step to design an automated approach in order to determine fake from genuine medicines without a need for trained chemists. The authors desired to use two well-known and widely used of machine learning methods to validate the results. For the purpose of this study RBF-SVM (Support Vector Machines with kernels) and KNN (K-Nearest Neighbor) were trained.

II. Machine Learning Models

A brief outline of the two machine learning techniques that was trained in this paper is presented in this section. They are k-Nearest Neighbor algorithm and Support Vector Machine which are briefly defined below.

A. *The KNN (k-nearest neighbor algorithm)*

KNN can be identified as an approach that is used to categorise substances based on a method for classifying objects based on contiguous training instances in the features universe [8]. The algorithm of k-nearest neighbor is assumed as non-complicated learning procedure which stores feature vectors and category labels of the training data. During the classification procedure, features that without labels is allocated to its nearest k neighbors category label. Characteristically the majority vote is the method that is used to classify the object, based on the labels of its k nearest neighbors. So for example if k is equal to 1, the object then classified to the class of the object nearest to 1. However, if there are just two classes, the value that represents k must be an odd integer. The main idea of this algorithm is to use Euclidean distance after representing the data to real numbers vectors as shown in the following equation:

$$d(a,b) = \|a-b\| = \sqrt{(a-b).(a-b)} = (\sum_{i=1}^{m}(a_i - b_i)^2))^{1/2} \quad (1)$$

where a and b are histograms in A = R^m

B. *SVM (Support Vector Machines)*

SVM is considered as one of the most effective algorithms for supervised twofold classification process for high dimensional data . This algorithm has been developed by Cortes and Vapnik (1995) and categorized as a linear algorithm. SVM is modeled as below:

$$h(a) = b + \sum_{n=1}^{N} y_i c_i K(a, a_i) \quad (2)$$

Where,

- h(a) represents the value of distance of the decision edge.
- b represents the value of the bias weight.
- c represents the coefficient that maximize the correct classification margin based on on the training set's elements.
- N represents the value of the number of features.
- K represents the value of the kernel function.
- a represents the features vector.

The SVM can use several kernels, such as the radial basis function, linear, or polynomial. The kernel function aims to divert the objects into a higher dimensional space as a proactive step in order to apply the classes' separation in non-linear space [9]. The SVM attempts were applied using three methods. the first is the linear SVM application. The second is SVM with radial basis function (RBF) kernel, which is used to make a nonlinear feature map (implicit). RBF works on adding a "bump" around each data point. The third is the implementation of the sequential minimal optimization algorithm (SMO) which was used to perform the training process for support vector classifier SVM [10].

III. Experiment and Material

Ten samples of Tenormin® containing 50 mg of active and inactive elements supplied by the central hospital for authentic drug samples and local drugstores for mixed samples in Almuthanna province. Seven tablets containing active elements of several featured manufactures and three counterfeit ones were supplied and analysed via XRF-Minipal2. The counterfeit samples where manufactured in China and India from different manufactures. As this research intended to define the elements of Tenormin 50mg so the result of the most commonly used excipients is analysed by XRF-Minipal2. The standard excipients were starch, lactose, talc, magnesium stearate, sodiccroscarmelose, ethylcellulose, microcristalina cellulose, sodium lauril sulphate and silicon dioxide. Tablets from each manufacture were finely powdered. Analyses were performed using an X-ray fluorescence minipal 2 from Philips®. Our proposed technique has two main phases which include feature components.

A. *XRF-Minipal2 Analyser*

The XRF-Minipal2 key component is its internal intelligence. It works based on a robust software. This software performs as good as other bench-top ED-XRF systems with their complicated filtering systems and multifaceted hardware configurations. This software also enhances the flexibility during the elemental analysis and fasten the process across the full spectrum of the periodic table. The XRF-Minipal2's distinctive software approach gives the capability of bench-top ED-XRF system to analyse virtually all types of elements with no special equipment or adjustments. The software is Windows®-based method that adds more unique features and facilities. The XRF-MiniPal2 is considered as a simple device when needs to be operated. It is also assumed as a stress-free as no extensive knowledge needed to operate analysis routine as the software does all tasks. The figures below show the XRF-Minipal2 system equipment, Figures 1 and 2 shown the off-on device status.

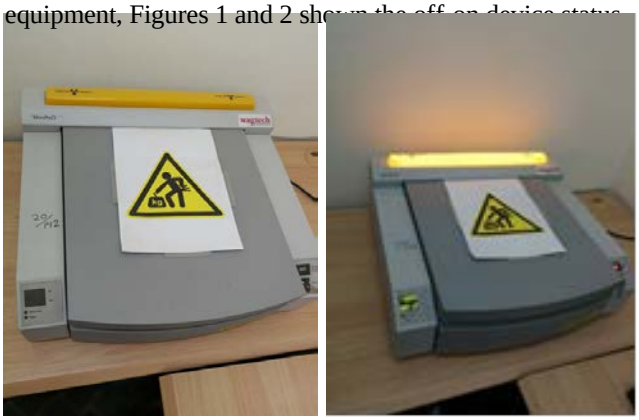

Fig 1&2: shows the XRF-Minipal2 device on Off/On status

While figure 3 shows the cubs where the elements that need to be placed in order to be tested. Two types of materials; solid on the form of powder and liquid material. Figure 4 shows the full XRF-Minipal2 system equipment which includes the X-ray fluorescent device and dedicated computer that capable to run the unique analyzing software which as shown Windows-based. While figure 5 shows the main software interface.

Fig 3: Shows the cubs that elements needs to be placed

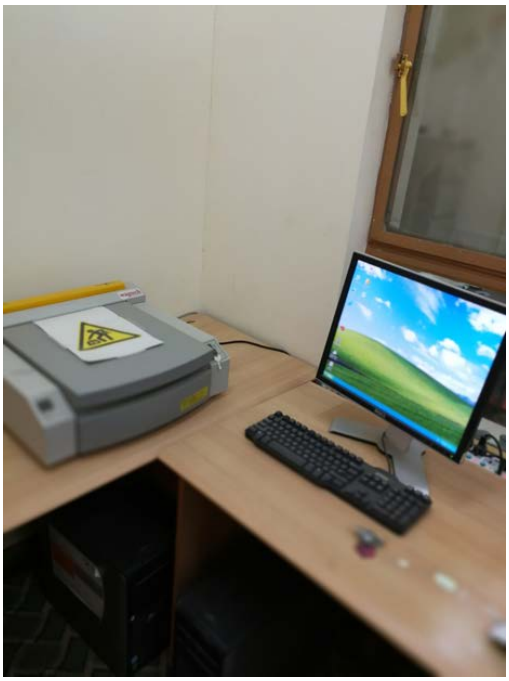

Fig4: full XRF-Minipal2 system equipment which include (XRF device + Software)

B. *Classifier*

In order to monitor the quality of pharmaceutical products and differentiate between the genuine and counterfeit medicines, there is a compulsory need to develop an automated dominant analytical techniques. For the purpose of this research a machine learning algorithm has been added as an extra layer for classification task. In order to validate the results the authors were employed two machine learning that work on non-linear concept. The first one is *k*-nearest neighbor classifier, where both variables of distance metric which called "nearest" and the number of objects that so-called "neighbors" can be modified. The Matlab environment was chosen to apply classification models. As we have two class labels " genuine" and "counterfeited", the Fine KNN classifier type was applied with the number of neighbors that set to =1. While other attributes such as prediction speed and memory usage are "Medium" due to the limited number of samples.

For the second classifier SVM-RBF (radial basis function) kernel, instead of the linear SVM was chosen. The kernel function aims to divert the objects into a higher dimensional space as a proactive step to apply the separation of classes in non-linear space. The SVM with RBF kernel depends on two parameters (C and γ) with the tune.svm () function. For this method the recommendations by [13] was followed, they set the same value for both parameter selection (C, γ) for all classifiers.

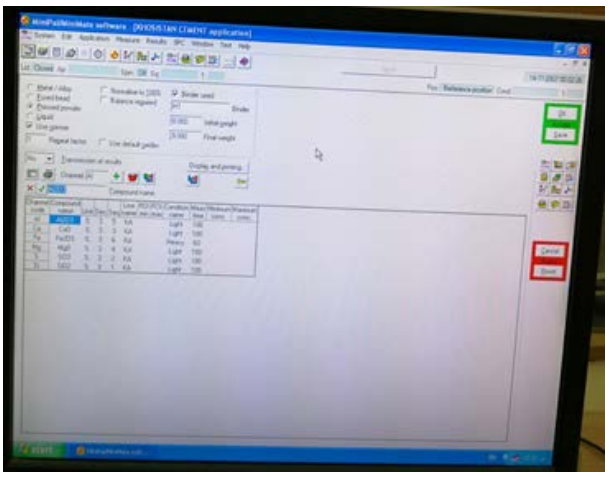

Fig 5 shows the main software interface

IV. RESULTS

Based on XRF-Minipal2 analysing system it was noted that among the set of counterfeit Tenormin® (3 cheap commercial samples) the inactive elements concentration in all tested samples is registered as much higher than the standard levels that registered for genuine samples (200%), while the active elements concentration levels are

remarkably lower than the standar The XRF-Minipal2 analysed data were subjected to be classified by two different machine learning algorithms.

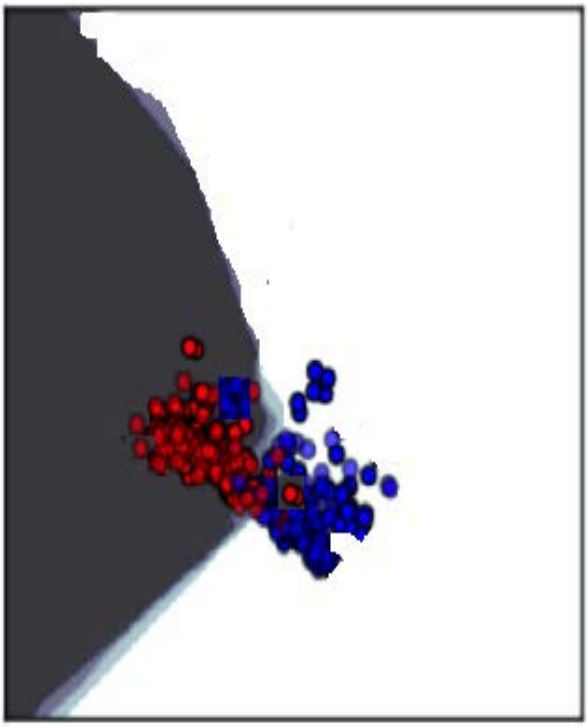

Fig 6: Visual illustration of KNN model representing the degree separations of genuine and counterfeited drugs.

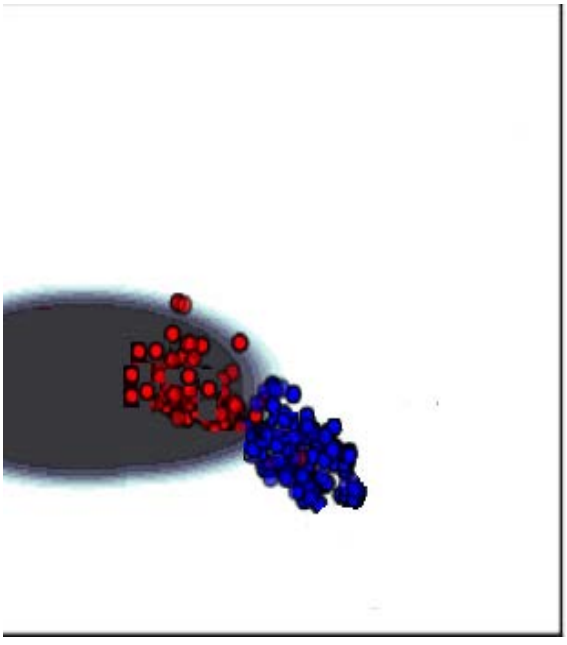

Fig 7: Visual illustration of RBF-SVM model representing the degree separations of genuine and counterfeited drugs.

The classification algorithms were trained using leave-one-out-cross-validation (LOOCV). This type of validation consumes a large number of computations but it is very accurate as the error rate based on a single instance. The process of LOOCV has iterated N (N=the number of samples) times and the results from N tested samples are calculated, then the results of all N samples are averaged to produce a single estimation. The Pseudocode for (LOOCV) is as below.

Leave_One_Out_Cross-_Validation (LOOCV) Procedure.
1. The data_set is divided into k examples.
2. One sample out of the k is left for validation purposes
3. The residual k-1 samples are used to train the model.
4. The steps from 1-3 is reiterated (n) times, using one trial set each time to validate the data as an validation sample.
5. The obtained results which are equal to (n) are then averaged to calculate the mean error (M_E) over each individual element through the cross-validation routines..

Fig 6 and 7 have shown a visual illustration of drug samples as genuine or counterfeited, by being subjected to supervised classification. Two algorithms were implemented; "K-Nearest_Neighbor" and "Radial_Basis_Function" which can be abbreviated as "RBF-SVM". The both models have clearly separated the two classes of items. Table 1 presents the overall outcomes of the supervised algorithms. The accuracy for the both supervised models has shown that RBF-SVM outperformed the KNN model. The results suggested that the models were able to generalize better as more patterns arose from more samples.

Table 2 Results of the performance of KNN and RBF-SVM

Classifier	**Accuracy %**
KNN	***91***
RBF-SVM	97

To clarify the above discussion Fig. 8, presents the confusion matrices which have 4 sectors, True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). The sector of **TP** indicates that a counterfeited medicine is properly detected as counterfeited. **TN** indicates that a genuine medicine is properly detected as genuine. However, **FP** indicates that a genuine medicine is improperly detected as counterfeited. Lastly, **FN** indicates that a counterfeited medicine is incorrectly detected as genuine. This paper's proposed technique that based on machine learning algorithms are designed to give the priority to detect counterfeit medicine which was represented by the **TP** (True

Positives). As this is our main goal, RBF-SVM outstanding the KNN model with (97.8%) and (9.9) for KNN. In Fig 8 the **TP**,**TN**, **FP** and **FN** are clearly depicted for both models.

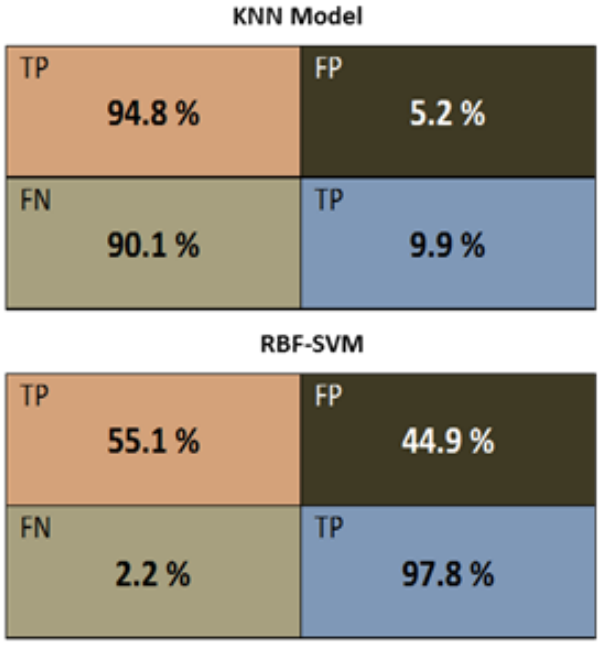

Fig 8: Confusion matrix for two models KNN and RBF-SVM

V. CONCLUSION

This paper presents XRF-Minipal2 elemental composition device that performs as fast screening technique in forensic investigations to determine genuine and counterfeited medicine. Tenormin 50 mg from different manufactures were analysed. The original fingerprints of genuine and counterfeited medicine were characterised. An extra machine learning analysing layer was proposed as a step for building-up an automated classifier. The analysed data that obtained from elemental analyser were used as an input for two supervised machine learning algorithms; RBF-SVM and KNN. Those two supervised learning models for classification task have been implemented to detect Tenormin drugs as either genuine or counterfeited based on their labels. The results revealed that our proposed technique is promising for detecting counterfeit medicine and capable to control it. For future work, more medicine samples will be used. Some potentials that this investigation trend could be taken further by implementing deep machine-learning (DML) models and improve the technique components. The authors intend to improve the performance by applying Latent semantic analysis component in order for dimensionality reduction and uncover latent correlations between features.

REFERENCES

[1] Alnajjar, A. (2017). Detection of Counterfeit Sildenafil Tablets Using Neutron Activation Analysis. [online] Nauss.edu.sa. Available at: http://repository.nauss.edu.sa/123456789/65083 [Accessed 14 Jun. 2018].

[2] Seear, Michael, D. Gandhi, R. Carr, A. Dayal, D. Raghavan, and N. Sharma. "The need for better data about counterfeit drugs in developing countries: a proposed standard research methodology tested in Chennai, India." Journal of clinical pharmacy and therapeutics 36, no. 4 (2011): 488-495.

[3] Dempsey, Adam, and Theodore Karasik. "MENA's Fake Pharma Conundrum." IndraStra Global 6 (2017): 3.

[4] Shiraiwa, Toshio, and Nobukatsu Fujino. "Theoretical Calculation of Fluorescent X-Ray Intensities in Fluorescent X-Ray Spectrochemical Analysis." Japanese Journal of Applied Physics 5, no. 10 (1966): 886.

[5] Cesareo, Roberto. X‐Ray Fluorescence Spectrometry. Wiley‐VCH Verlag GmbH & Co. KGaA, 2010.

[6] Van Grieken, Rene, and Andrzej Markowicz, eds. Handbook of X-ray Spectrometry. CRC Press, 2001. pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Beckhoff, Burkhard, Birgit Kanngießer, Norbert Langhoff, Reiner Wedell, and Helmut Wolff, eds. Handbook of practical X-ray fluorescence analysis. Springer Science & Business Media, 2007.

[8] AlSallal, Muna, Rahat Iqbal, Vasile Palade, Saad Amin, and Victor Chang. "An integrated approach for intrinsic plagiarism detection." Future Generation Computer Systems (2017).

[9] Alsallal, Muna. "A Machine Learning Approach for Plagiarism Detection." PhD diss., Coventry University, 2016.

[10] AlSallal, Muna, Rahat Iqbal, Saad Amin, Anne James, and Vasile Palade. "An Integrated Machine Learning Approach for Extrinsic Plagiarism Detection." In Developments in eSystems Engineering (DeSE), 2016 9th International Conference on, pp. 203-208. IEEE, 2016.

[11] W. Romão, M.F. Franco, M.I.M.S. Bueno, M.N. Eberlin, M.-A. De Paoli, Analysing metals in bottle-grade poly(ethylene terephthalate) by X-ray fluorescence spectrometry, J. Appl. Polym. Sci. 117 (2010) 2993–3000.

[12] W. Romão, P.M. Lalli, M.F. Franco, G. Sanvido, N.V. Schwab, R. Lanaro, J.L. Costa, B.D. Sabino, M.I.M.S. Bueno, G. de Sa, R.J. Daroda, V. Souza, M.N. Eberlin, Chemical profile of meta-chlorophenylpiperazine (m-CPP) in ecstasy tablets by easy ambient sonic-spray ionization, X-ray fluorescence, ion mobility mass spectrometry, and NMR, Anal. Bioanal. Chem. 400 (2011) 2685–2686

[13] Zheng, R., Li, J., Chen, H. and Huang, Z., (2006) 'A Framework for Authorship Identification Of Online Messages: Writing‐Style Features And Classification Techniques'. Journal of the American Society for Information Science and Technology, 57 (3), 378-393.