Medical Formulation Recognition (MFR) using Deep Feature Learning and One Class SVM

Omar Kawi
University of Sunderland
& Rokshaw Laboratories
Sunderland, UK
omar.alkawi@sunderland.ac.uk

Kathy Clawson
Faculty of Technology
University of Sunderland
Sunderland,UK
kathy.clawson@sunderland.ac.uk

Paul Dunn
Rokshaw Laboratories
Sunderland,UK
paul@rokshaw.co.uk

Daniel Knight
University of Sunderland &
Rokshaw Laboratories
Sunderland,UK
daniel@rokshaw.co.uk

Jonathan Hodgson
Rokshaw Laboratories
Sunderland, UK
jonathan@Rokshaw.co.uk

Yonghong Peng
Faculty of Technology
University of Sunderland
Sunderland,UK
yonghong.peng@sunderland.ac.uk

***Abstract*—Specials medications are personalized formulations manufactured on demand for patients with unique prescription requirements and constitute an essential component of patient treatment. Specials are becoming increasingly in demand due to the need for personalized and precision medicine. The timely provision of optimal personalized medicine, however, is challenging, subject to strict regulatory processes, and is expert intensive. In this paper, we propose a new medical formulation engine (MFE) that performs semantic search across multiple disparate formulations archives to enable data driven formulation intelligence. We develop a new platform for medical formulations recognition (MFR) that curates a new dataset comprising formulations and non-formulations (clinical) text and uses a novel pipeline encompassing deep feature extraction and one-class support vector machine learning. The proposed MFR framework demonstrates promising performance and can be used as a benchmark for future research in formulations recognition.**

Keywords—Text Recognition, NLP, Deep Learning, One-Class Learning, Support Vector Machine

I. Introduction

The majority of medicines prescribed to adults are licensed products with clearly defined usage and will have undergone rigorous regulatory procedures to evidence drug efficacy, safety, administration, indications and shelf-life. However, the use of unlicensed, off-label, and personalized medication is also common, especially within certain patient demographics such as neonates and the elderly [1]. It has been stated that 93% of neonates in intensive care will receive at least one unlicensed or off-label medicine [2]. Specials medications are formulations which are bespoke manufactured for patients with unique prescription requirements, specifically for individuals who clinically require something that is different from the standard licensed format [1]. This may be because: they are a baby/small child or elderly and require a different strength or format, they are allergic to ingredients, have swallowing difficulties [3], or other complexities; the drug is new and / or there is not enough demand; or because there are supply issues with a licensed product [4]. According to the Association of Pharmaceutical Specials Manufacturers, unlicensed and off-label medication represents 1% of total prescriptions and constitutes more than 75 000 formulations per annum. In 2017 the NHS drug spend was £9.17 billion, and £77.5 million constituted specials [1].

The individualized nature of specials production, coupled with requirements for strict quality control, constitutes a substantial burden for healthcare professionals. Such practice is a necessary part of product development and should be informed using the best available evidence. Furthermore, it is necessary to provide efficient and timely drug provision in order to avoid practical problems such as delays in treatment [4]. Access to existing bodies of clinical knowledge is an important factor and is reflected in the sheer volume of formulation studies which have been published since the 1960s [5,6] Such studies are typically long-term, analyze drug content and drug degradation, and consider a variety of environmental factors (such as temperature, light and pH).

Despite the availability of vast digital archives of clinical publications including formulation studies, archive search is non-trivial. Formulation development and validation is an inherently manual process which involves expert input. There

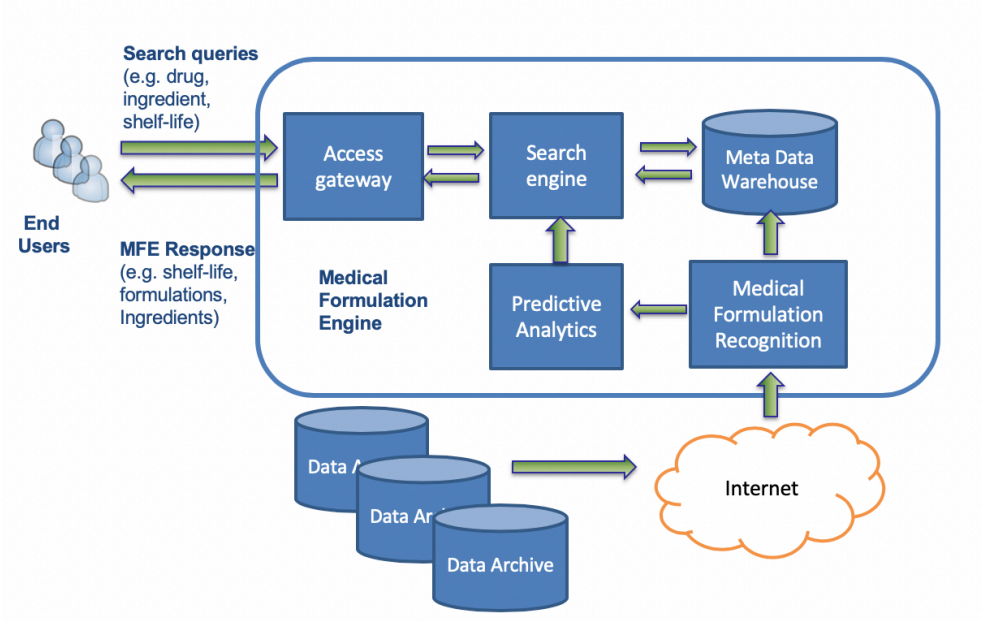

Fig. 1. Medical Formulation Engine Architecture

exists a high similarity between formulations text and other medical studies, such as clinical trials, and information is typically stored across multiple, disparate archives of data (both proprietary and non-proprietary). These factors result in time inefficiencies and are costly for the industry.

There exists a clear opportunity to exploit current trends in natural language processing and machine learning, such as deep learning, to assist with specials production. In this paper we propose the development of a Medical Formulation Engine (MFE), which enables user-defined search of multiple data archives and utilizes state-of-the-art machine learning methods including natural language processing (deep feature extraction coupled with one-class learning) for automatic recognition, retrieval, and creation of pharmaceutical formulations (Fig. 1.). Our overriding objective is to enhance existing operating procedures and facilitate data – driven knowledge creation within the specials manufacturing market, to enhance personalized medicine, enable intelligent formulation search and facilitate further work on predictive analytics. Specifically, this work addresses the following challenges:

- There is no single public dataset available which represents medicine formulation studies.
- Public sources of formulations data are typically heterogeneous and disparate.
- Practitioners will access multiple sources manually when building specials formulations and will utilize both internal and external data.
- There exists a high semantic similarity between formulations data and other clinical sources- both expert search and automated recognition are non-trivial.

The remainder of this document is structured as follows. Our methodology for medical formulations recognition is presented in Section 2. An experimental overview is offered in Section 3. Results and discussion are offered in Section 4, and Section 5 describes conclusions and future work.

II. Method

A. System Overview

This study seeks to evaluate the accuracy with which pharmaceutical formulations text may be recognized from other similar clinical manuscripts. Our framework for Medical Formulation Recognition (MFR) is presented in Fig. 2. After acquisition, data is transformed into deep feature vectors using Universal Sentence Encoding (USE) [7] and resultant feature-space representations are reduced via Principal Component Analysis (PCA) [8]. Due to problems associated with learning imbalanced data (e.g. the existence of a skewed class distribution, and the potential for over-training in favor of the majority class), we regard pharmaceutical formulation recognition as a one-class problem, where non-formulations

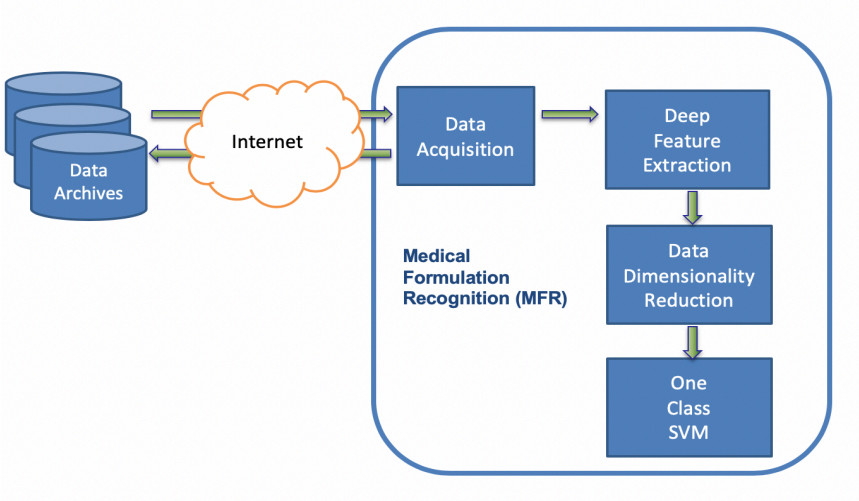

Fig. 2. MFR Methodology.

(clinical) text is defined as an outlier. To achieve this, *n* principal components of USE features are used as inputs for one-class support vector machine (SVM) learning. One-class SVM learning [9] is an approach that has been successfully applied to overcome data imbalance across a variety of application areas, including fraud detection [10], pronunciation verification [11], and cancer diagnosis [12].

A. Data Acquisition

Journal abstracts were digitally collected from the following sources:

- PubMed (life sciences & biomedical) repository using the BioPython Web API [13]
- Trissel's online archive of compounded formulations [14], via (keyword-based) web crawl and abstract extraction.

For PubMed inputs, in order to retrieve articles corresponding to unlicensed medicines formulations, we made API calls using queries which were co-produced with a team of experts offering domain insight. After search query execution using the BioPython Web API, returned abstracts were inspected and manually labelled by experts, with each abstract classified as corresponding to a formulation or non-formulation study. Example queries are illustrated in Fig. 3. For SVM learning, only formulation abstracts were retained.

Our complete data set comprises 968 abstracts, of which 882 and 86 constitute formulations and non-formulations, respectively. All abstracts are clinical in nature. Abstract lengths range between 2 – 20 sentences, with a mean length of 9.23 and standard deviation of 3.18. Data was stored as unstructured raw text and used directly as inputs for feature extraction. 308 formulations abstracts were sourced from Trissels [8], and the remainder of abstracts were the result of PubMed search.

```
queryList =                          ['Stability aqueous solution',
                                          'compounded formulation',
                                 'compounded formulation oral suspension',
                                        'extemporaneous stability',
                                         'oral solution stability',
                                    'stability ointment  90 days',
                                              'Stability ora plus']

for q in queryList
    results = search(q)

def search(query):
    handle = Entrez.esearch(db='pubmed',
                            sort='relevance',
                            retmax='10000',
                            retmode='xml',
                            term='query')
    results = Entrez.read(handle)
    return results
```

Fig. 3. Example WebAPI Queries

B. Deep Feature Extraction

In this research, feature extraction is regarded as a transfer learning task and implemented using Universal Sentence Encoding (USE). USE has been trained and optimized for greater-than-word-length NLP activities, and can take as input sentences, phrases or short paragraphs. The model is trained with a Deep Averaging Network (DAN) encoder [15] and does not require text pre-processing [7]. It encodes text into a fixed-dimensional (512 features) embedding string representations [7] and has been successfully applied across a variety of NLP tasks including text mining, document classification, clustering, and semantic similarity. USE achieves good performance with minimal amounts of training data [7], which makes it appropriate in scenarios where large training sets are not available.

C. *Dataset Dimensionality Reduction*

We perform data dimensionality reduction using Principal Components Analysis (PCA). PCA, alternatively known as the Hotelling Transform, is an unsupervised linear transform performed by calculating the eigenvectors of a dataset's covariance matrix and projecting resultant data onto a new coordinate system where the data is mapped in decreasing order of variance [8]. By retaining only those components with higher variance, we reduce UCE vectors into a smaller set of variables, aim to decrease model complexity and associated training times, and avoid overfitting.

D. *One Class Support Vector Machine (OCSVM)*

OCSVM is an unsupervised learning technique for outlier detection which was first introduced by Schölkopf et al [9]. OCSVM modelling is distinct from multi-class supervised SVM learning, given that inputs belong to a single class of data and can therefore technically be regarded as unlabeled [16]. Within the literature, [17] utilize OCSVM for gathering rich data from medical subject headings (MeSH). [18] utilize one class SVM for document classification, and similar approaches have also been applied on image processing applications such as detecting chinese calligraphy style differences [19]. However, the application of OCSVM learning for formulations recognition has been previously unexplored.

The one-class SVM learning problem is framed as: Given a dataset with feature space probability distribution *P*, find a "simple" subset S of the feature space such that the probability that a test point from P lies outside S is bounded by some a-priori specified value [9]. To generate the boundary and separate the dataset from the origin [9], we solve:

$$\min_{w\epsilon F,\varepsilon\epsilon\mathbb{R}^{\ell},\rho\epsilon\mathbb{R}} \quad \frac{1}{2}\|w\|^2 + \frac{1}{\upsilon\ell}\sum_i \varepsilon_i - \rho \quad (1)$$

$$\text{subject to} \quad (w.\Phi(x_i)) \geq \rho - \varepsilon_i,\ \varepsilon_i \geq 0\ \forall i$$

Where x_i is the training data points, ℓ the number of observations, Φ is feature map, and υ represents the upper bound on the fraction of outliers and lower bound on the fraction of support vectors. Moreover, ρ and w are decision variables define the classifiers, ε_i non-zero slack variables penalized in the object function [20][9].

Equation (2) will be used instead of (1) as its "common to solve the Lagrange dual":

$$\min_{\alpha} \quad \frac{1}{2}\sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) \quad (2)$$

$$\text{subject to} \quad 0 \leq \alpha_i \leq \frac{1}{\upsilon\ell},\ \sum_i \alpha_i = 1$$

where k represents the kernel function, where the kernel function is transforming the data input into specific form. In one class SVM there are many types of the kernel functions such as; linear classifier, polynomial, radial basis function (RBF) and sigmoid. In this work we will use only RBF kernel function, where k is:

$$k\,(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (3)$$

Where $\|x - y\|^2$ is the Euclidean distance between two data points, and σ is free parameter of the kernel function.

III. Experimental Overview

After data acquisition using the approach defined in Section II, we partition abstracts into training and test samples. It has been previously reported that OCSVM operates better when there are no or less anomalies in the training data [10, 21]. For this reason, we exclude non-formulation abstracts (negative samples) from model training and partition our formulations data into 90% training and 10% test sets. The formulations test set (89 samples) is subsequently combined with non-formulation data (86 samples) for final system evaluation. Our full data partitioning protocol is illustrated in Table 1.

An example of formulations text is illustrated in Fig. 4. USE takes as input lowercase strings. Due to unique characteristics of our data, we do not perform any additional pre-processing, and encode data at the paragraph level. Formulation abstracts are typically short and contain domain specific numerical data and special characters. We wish to maintain domain specific content and investigate the accuracy achievable using simple models which are trained at the paragraph level. After USE, we reduce each document's 1*512 feature vector to a 1*n vector of principal components, with n = 30 selected after empirical investigation of component variance (Fig. 5.). Specifically, we retain only those components required to maintain 80% of total dataset variance.

TABLE I. Training & Test Set Paradigm

Description	Experimental Setup		
	Training	***Testing***	***Total***
Formulation	793	89	882
Non-Formualtion	0	86	86
Total	793	175	968

> formulation a stayed physicochemical and microbiologically stable at refrigerated (4°c) conditions during at least 150 days and it only stayed stable during 14 days at 25°c. formulation b was stayed physicochemical and microbiologically stable at refrigerated (4°c) conditions at least 90 days, but it is not recommended to store at 25°c for more than 1 day.

Fig. 4. Example Formulations Text

Our OCSVM is trained using an RBF kernel. Given that parameter optimization is a significant issue for OCSVM [22], we evaluate performance across a variety of outlier fractions (v) and *gamma* values using the following metrics:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (4)$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (5)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (6)$$

and

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Samples} * 100 \quad (7)$$

Our system is implemented in python using Scikit-learn and TensorFlow packages [23, 24], and executed within a Windows environment (i7 processor, 16 GB RAM).

IV. Results & Discussion

Mean medical formulation recognition accuracy across the full range of v and *gamma* parameterizations was 0.752, with precision = 0.83, recall = 0.66, and f1 = 0.722. Those individual SVM setups achieving maximum performance for each metric are summarized in Table II. Across all experiments, maximum classification accuracy of 0.817 was attained (F1 score = 0.832).

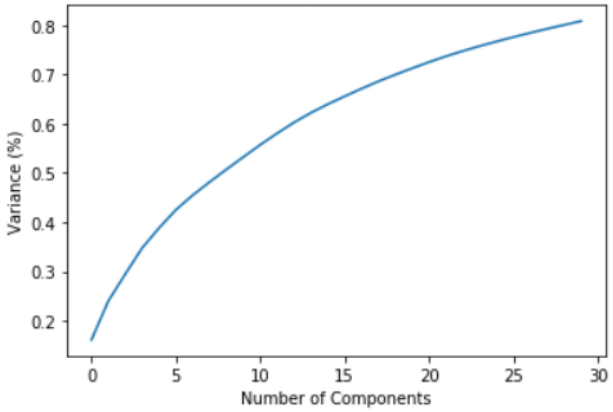

Fig. 5. PC Variances

TABLE II. Summary of Classifier Performance

v	*gamma*	*Accuracy*	*Precision*	*Recall*	*F1*
0.0100	5.3000	**0.817**	0.782	0.888	**0.832**
0.4200	8.9000	0.731	**0.977**	0.483	0.647
0.0100	1.3000	0.789	0.706	**1.000**	0.828

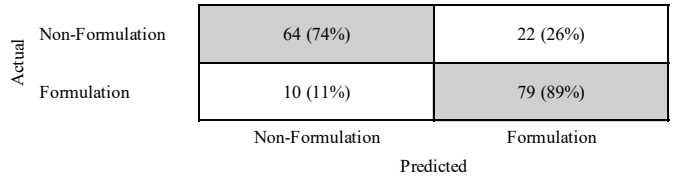

Fig. 6. Confusion Matrix, Gamma = 5.3, *v = 0.01*

The confusion matrix for v = 0.01 and gamma = 5.3 is provided in Fig. 6. It can be seen from Fig 6. that 89% of formulations were correctly recognized, and 26% of non-formulations were misclassified. Full inspection of accuracy as a function of OCSVM parameterization (Fig. 7) illustrates: accuracy > 0.74 where v < 0.25; and accuracy > 0.78 with mid-range *gamma* values. Similarly, F1 scores are maximized when $v < 0.2$ (Fig. 8). It apparent from Table II that there exists a trade-off between classification accuracy and system precision and recall. Where accuracy = 0.817, precision is 0.197 less than the maximum achievable (max precision = 0.977).

This observation demonstrates the importance of considering a range of performance metrics when working with imbalanced datasets, and of evaluating classifier sensitivity to parameterization. We may increase mean precision (Fig. 9), but this is at the expense of true positive prediction. Analysis of mean medical formulation recognition (across all *gamma* values) as a function of v (outlier fraction), as illustrated in Figure 9, further highlights this. Specifically, there exists an inverse relationship between precision and our other performance metrics. Increased precision is at the expense of recall. Where recall = 1, 43% of non-formulation abstracts are incorrectly recognized (Figure 11).

V. Conclusions

Medical formulations recognition (MFR) raises a promising and challenging task and offers excellent opportunities for further research in this area. This paper presents a new MFR dataset and evaluates the performance of the application of deep feature extraction and one class support vector machine learning for formulation recognition. Initial results demonstrate the promising performance of our proposed approach. Mean recognition across all SVM parameterizations is 0.752 and through OCSVM parameterization we can achieve accuracy of 0.82, with 0.78 precision and 0.88 recall. We propose that our dataset and methodology constitute a new benchmark facilitating further research in this area.

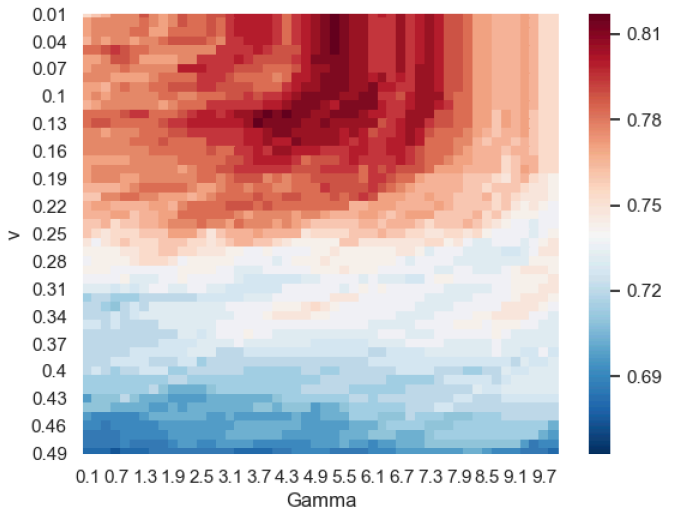

Fig. 7. Classification Accuracy as a Function of Parameterization

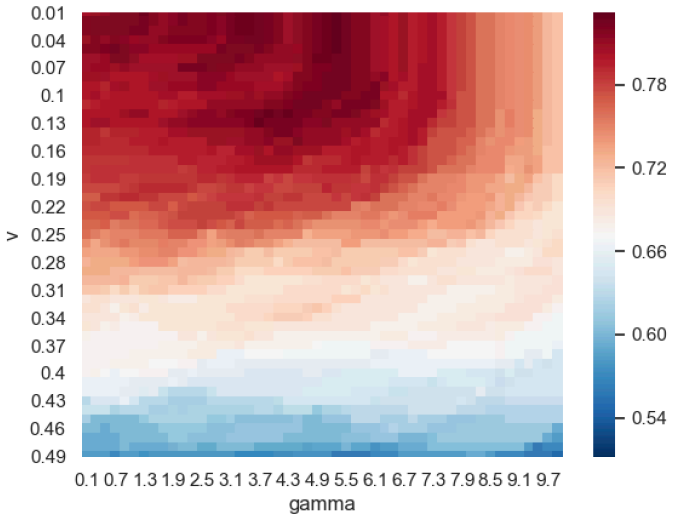

Fig. 8. F1 Score as a Function of Parameterization

When applied to MFR, the OCSVM demonstrates a clear sensitivity to parameterization. Future work could therefore focus on enabling robust, generalizable learning. This can be achieved through adoption of more sophisticated feature extraction and representation methods, comparison of multiple kernels for learning, and further fine-tuning. The USE approach utilized constituted full transfer learning, with no fine-tuning, implemented at paragraph level. A comparative investigation of sentence- versus paragraph- level feature extraction, with fine-tuning, is therefore desirable. Furthermore, there exists the opportunity to integrate additional approaches for feature extraction and representation, and machine learning models (for example novel approaches to MFR incorporating deep convolutional neural networks).

Our goal is to have an automatic medical formulation engine (MFE) that can efficiently retrieve formulations from published archives based on user requests and create the formulation for the special requests. The output of the MFR will be used inside the Medicine Formulation Engine (MFE) to enable rapid formulation and production of novel and bespoke medications. MFE will assist in finding new formulations, where finding these formulations currently takes significant manual time and effort. This research and the MFE search engine will offer time and cost efficiencies to pharmacists searching through published formulation and stability studies, with the aim of reducing overall product development times and enhancing patient outcomes through timely provision of personalized care.

ACKNOWLEDGMENT

We acknowledge the support of Innovate UK under grant number 11470.

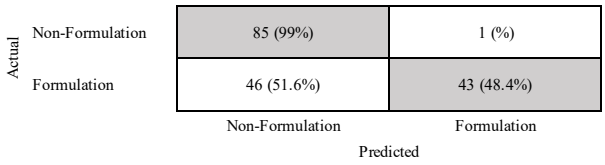

Fig. 9. Confusion Matrix, Gamma = 8.9, v = 0.42

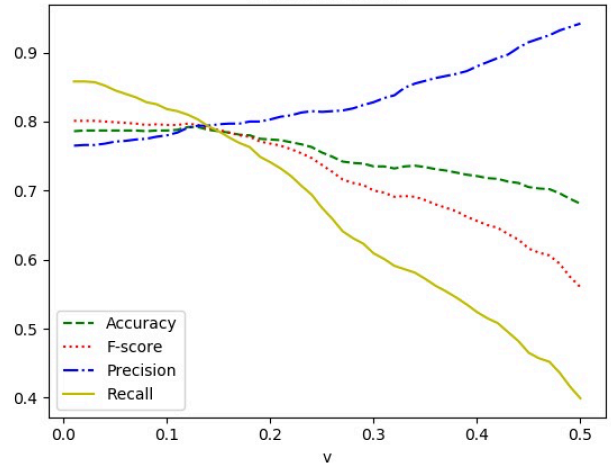

Fig. 10. Mean Performance as a Function of Outlier Fraction (v)

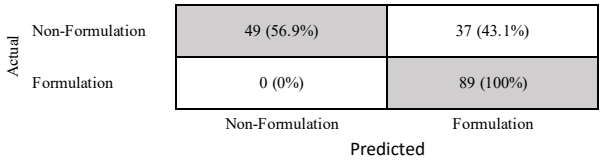

Fig. 11. Confusion Matrix, Gamma = 1.3, v = 0.01

REFERENCES

1- G. Donovan, L. Parkin, L. Brierley-Jones and S. Wilkes, "Unlicensed medicines use: a UK guideline analysis using AGREE II", *International Journal of Pharmacy Practice*, vol. 26, no. 6, pp. 515-525, 2018. Available: 10.1111/ijpp.12436 .

2- S. Conroy and J. McIntyre, "The use of unlicensed and off-label medicines in the neonate", *Seminars in Fetal and Neonatal Medicine*, vol. 10, no. 2, pp. 115-122, 2005. Available: 10.1016/j.siny.2004.11.003 .

3- Lowey, Andrew, and Mark Jackson. "How to ensure the quality and safety of unlicensed oral medicines." *Acute pain* 10 (2019): 00.

4- R. Griffith, "Unlicensed medicines", *British Journal of Nursing*, vol. 28, no. 17, pp. 1154-1155, 2019. Available: 10.12968/bjon.2019.28.17.1154 .

5- Sadrieh, Nakissa, James Brower, Lawrence Yu, William Doub, Arthur Straughn, Stella Machado, Frank Pelsor et al. "Stability, dose uniformity, and palatability of three counterterrorism drugs—human subject and electronic tongue studies." *Pharmaceutical research* 22, no. 10 (2005): 1747-1756.

6- [V. Chédru-Legros et al., "In Vitro Stability of Fortified Ophthalmic Antibiotics Stored at −20°C for 6 Months", *Cornea*, vol. 29, no. 7, pp. 807-811, 2010. Available: 10.1097/ico.0b013e3181c32573 .

7- Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant et al. "Universal sentence encoder." *arXiv preprint arXiv:1803.11175* (2018).

8- Smith, Lindsay I. *A tutorial on principal components analysis*. 2002.

9- Schölkopf, Bernhard, Robert C. Williamson, Alex J. Smola, John Shawe-Taylor, and John C. Platt. "Support vector method for novelty detection." *Advances in neural information processing systems*, pp. 582-588. 2000.

10- P. Zheng, S. Yuan, X. Wu, J. Li and A. Lu, "One-Class Adversarial Nets for Fraud Detection", *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 33, pp. 1286-1293, 2019. Available: 10.1609/aaai.v33i01.33011286 .

11- M. Shahin, J. Ji and B. Ahmed, "One-Class SVMs Based Pronunciation Verification Approach", *2018 24th International Conference on Pattern Recognition (ICPR),* 2018. Available: 10.1109/icpr.2018.8545687.

12- Xu, Haifeng. "An Evaluation of One Class Classifier on Gene Expression Data." *PhD diss., Tampere University,* 2019.

13- P. Cock et al., "Biopython: freely available Python tools for computational molecular biology and bioinformatics", *Bioinformatics, vol. 25, no. 11, pp. 1422-1423, 2009*. Available: 10.1093/bioinformatics/btp163.

14- L. Trissel, Trissel's stability of compounded formulations. Washington, D.C.: *American Pharmacists Association*, 2009.

15- Iyyer, Mohit, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. "Deep unordered composition rivals syntactic methods for text classification." *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers), pp. 1681-1691. 2015.

16- V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection", ACM Computing Surveys, vol. 41, no. 3, pp. 1-58, 2009. Available: 10.1145/1541880.1541882.

17- J. Cha, J. Kim and S. Park, "GRiD: Gathering rich data from PubMed using one-class SVM", *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC),* 2016. Available: 10.1109/smc.2016.7844911.

18- Manevitz, Larry M., and Malik Yousef. "One-class SVMs for document classification." *Journal of machine Learning research* 2, no. Dec (2001): 139-154.

19- Z. Jiulong, G. Luming, Y. Su, S. Xudong and L. Xiaoshan, "Detecting Chinese calligraphy style consistency by deep learning and one-class SVM", *2017 2nd International Conference on Image, Vision and Computing (ICIVC)*, 2017. Available: 10.1109/icivc.2017.7984523

20- J. Fleming, X. Yan and R. Lot, "Fitting Cornering Speed Models with One-Class Support Vector Machines", *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019. Available: 10.1109/ivs.2019.8814061.

21- M. Amer, M. Goldstein and S. Abdennadher, "Enhancing one-class support vector machines for unsupervised anomaly detection", *Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description - ODD '13*, 2013. Available: 10.1145/2500853.2500857.

22- L. Zhuang and H. Dai, "Parameter Optimization of Kernel-Based One-Class Classifier on Imbalance Text Learning", *Lecture Notes in Computer Science*, pp. 434-443, 2006. Available: 10.1007/978-3-540-36668-3_47 .

23- TensorFlow Hub", TensorFlow, 2020. [Online]. Available: https://www.tensorflow.org/hub/. [Accessed: 28- Jan- 2020].

24- Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems." *arXiv preprint arXiv:1603.04467* (2016).