Computer Methods and Programs in Biomedicine 243 (2024) 107917

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Comparison of machine learning methods in the early identification of vasculitides, myositides and glomerulonephritides

Rasmus Ryyppö [a,c,*], Sergei Häyrynen [c], Henry Joutsijoki [a], Martti Juhola [a], Mikko Seppänen [b]

[a] Faculty of Information Technology and Communication Sciences, Tampere University, Kanslerinrinne 1, Tampere 33014, Finland
[b] Rare Disease Center and Pediatric Research Center, New Children's Hospital, University of Helsinki and HUS Helsinki University Hospital, Helsinki, Finland
[c] Tietoevry Ltd, Espoo, Finland

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Residual neural network (ResNet)
Inception model
XGBoost
Rare diseases

ABSTRACT

Background: Rare disease diagnoses are often delayed by years, including multiple doctor visits, and potential imprecise or incorrect diagnoses before receiving the correct one. Machine learning could solve this problem by flagging potential patients that doctors should examine more closely.
Methods: Making the prediction situation as close as possible to real situation, we tested different masking sizes. In the masking phase, data was removed, and it was applied to all data points following the first rare disease diagnosis, including the day when the diagnosis was received, and in addition applied to selected number of days before initial diagnosis. Performance of machine learning models were compared with positive predictive value (PPV), negative predictive value (NPV), prevalence PPV (pPPV), prevalence NPV (pNPV), accuracy (ACC) and area under the receiver operation characteristics curve (AUC).
Results: XGBoost had PPVs over 90 % in all masking settings, and InceptionVasGloMyotides had most of the PPVs over 90 %, but not as consistently. When the prevalence of the diseases was considered XGBoost achieved highest value of 8.8 % in binary classification with 30 days masking and InceptionVasGloMyotides achieved the best value of 6 % in the binary classification as well, but with 2160 days and 4320 days masking. ACC were varying between 89 % and 98 % with XGBoost and InceptionVasGloMyotides having variation between 79 % and 94 %. AUC on the other hand varied between 72.6 % and 94.5 % with InceptionVasGloMyotides and for XGBoost it varied between 69.9 % and 96.4 %.
Conclusions: XGBoost and InceptionVasGloMyotides could successfully predict rare diseases for patients at least 30 days prior to initial rare disease diagnose. In addition, we managed to build performative custom deep learning model.

1. Introduction

Classification tasks with *machine learning* (ML) are quite common in medical domain and their difficulty varies between applications. Problems arise when classification is done with partial data, and in the case of identification of *rare diseases* (RD) it means, that we are not using all the data available. RDs are difficult to detect, and diagnosis is often delayed which makes the classification task challenging. Early identification with ML has yielded quite good results earlier with dementia research by So et al. [1], and when researching early identification of diseases, we need to examine the question of how early it is possible to identify. This is crucial, because in some cases early identification is not early enough for patients. Early identification means also that we will need to work with partial data as it cannot be defined as early identification if we are using all the possible data including the disease diagnoses. Shen et al. [2] similarly studied accelerated RD diagnosis with a combination of ML and recommender systems by *collaborative filtering* (CF). They achieved promising results by using *natural language processing* (NLP) and CF with *Tanimoto coefficient similarity* (TANI) and *k-nearest neighbor* (KNN) algorithm. Despite that, CF has weaknesses i.e., sparse data and scalability. Since RD data is by definition sparse and there is a future need for scalable models that performs well with current diseases and can be expanded with other diseases, we decided to combine three somewhat related inflammatory disease groups not only into disease specific but also into a binary model. Binary model studies the likelihood of an individual to have any of the studied vasculitides, myositides and

* Corresponding author at: Faculty of Information Technology and Communication Sciences, Tampere University, Kanslerinrinne 1, Tampere 33014, Finland.
E-mail addresses: rasmus.ryyppo@tuni.fi, rasmus.ryyppo@tietoevry.com (R. Ryyppö).

https://doi.org/10.1016/j.cmpb.2023.107917
Received 15 May 2023; Received in revised form 29 September 2023; Accepted 5 November 2023
Available online 7 November 2023

glomerulonephritides diseases.

RDs are commonly defined in Europe as diseases with population prevalence less than 5 individuals in 10 000, and they are frequently difficult to diagnose, severe, systemic or one organ diseases. They commonly lead to so-called diagnostic odysseys with multiple evaluations, imaging studies and laboratory tests. In Australian adults, about 21 % of respondents informed that they had to wait the diagnosis of a RD for 1–5 years, 22 % waited for 5–10 years and about 10 % had to wait for correct diagnosis more than 20 years. About 66 % underwent three or more doctor visits [3]. In an Australian survey on children, parents reported that 42 % respondents had to visit 3–5, 17 % 6–10, and 11 % more than 10 different physicians. Of respondents, 60 % reported that after symptom onset, correct diagnosis was achieved within one year, 32 % after 1–3 years, and 8 % for more than 3 years [4]. Similar results have been reported from the United States of America where on average receiving diagnosis took 7.6 years, while in the United Kingdom this took on average 5.6 years. During the diagnostic process, patients experienced 8 doctor visits, and received 2–3 misdiagnoses [5]. Difficulties in diagnosing RD result in delayed and inadequate or even harmful clinical management. Shortening and ending such odysseys could potentially result in clinical, psychosocial, and economic benefits to patients, their families, healthcare, and society [6,7].

Vasculitides, myositides and glomerulonephritides, for most part non-familial inflammatory diseases affecting muscles, vessels, and kidneys, belong to RD. Common symptoms for myositides are muscle weakness and raised skeletal muscle enzymes. There are disease subsets for myositides, which are polymyositis (PM), sporadic inclusion body myositis (sIBM), dermatomyositis (DM) and immune-mediated necrotizing myopathy [8]. Prevalence rates for PM and DM ranges between 1 and 9 in 100,000 and IBM is rarer and prevalence for it ranges 1–9 in 1, 000,000 [9–11]. Vasculitides' non-specific symptoms can be fever, weight loss and myalgia, and in addition there are specific symptoms or combination of symptoms that are specific for different subgroups. As subgroups there are large vessel vasculitis (LVV), medium vessel vasculitis (MVV) and small vessel vasculitis (SVV) [12]. Prevalence average of vasculitides is 1–9 in 100,000 people [13]. Glomerulonephritides' common symptoms are fluid retention and hypertension, but there are some non-specific symptoms which are similar with vasculitides such as fever and weight loss. There are few different etiological subgroups such as immune-complex glomerulonephritides and pausi-immune glomerulonephritides [12]. Prevalence of the glomerulonephritides is 1.6 in 100,000 people, but this varies between countries [14].

In a pre-study assessment in *Helsinki University Hospital* (HUS), solely providing highly specialized tertiary care to over 1.6 million inhabitants, these diseases appeared to be the most common RD groups with significant delay in reaching the diagnosis. In addition, during their disease courses, demand for resource-intensive supportive therapies increased significantly. Research of this magnitude has not been done earlier when comparing the amount of data that can be used and having an objective of early identification of these specific diseases.

In healthcare systems with *electronic patient records* (EPR), ML and *diagnosis decision support systems* (DDSS) i.e., *Rare Disease Auxiliar Diagnosis* (RDAD) system introduced by Jia et al. [15] could potentially offer healthcare professionals an invaluable tool for early identification of RD. While using any ML applications in the healthcare is still uncommon, interest towards DDSS and other ML applications is increasing as capabilities of ML and *artificial intelligence* (AI) evolve. *Residual neural networks* (ResNet) were introduced by He et al. [16]. ResNet with the InceptionTime model showed very good results in the field of image classification problems and time series problems [17]. XGBoost is a state-of-the-art tree boosting method which has shown its capabilities with sparse data [18].

At an earlier stage, we developed InceptionVasGloMyotides model and transformed our dataset to be compatible with XGBoost. We established that the InceptionVasGloMyotides model was competitive against XGBoost in the early identification of RDs, especially in longer prediction periods. In addition to these, we did test ResNet and InceptionTime models, but their resolution did not perform at sufficient levels, which was presumable caused by the sparseness of the data and pooling method [19]. Here, we novelly compare XBoost with an InceptionVasGloMyotides model customized for RD diagnostics.

In Section 2, we describe InceptionVasGloMyotides and XGBoost. Then in Section 3, we will define the experimental setup. This includes description of data, preprocessing and used performance measures. Section 4 covers the results and in Section 5 we compare our paper against RD detection paper and paper with similar dataset format as ours. Finally, Section 6 concludes our paper.

2. Methods

2.1. InceptionVasGloMyotides

InceptionVasGloMyotides is Inception type ResNet which is a type of *convolution neural network* (CNN) model. Difference between ResNets and conventional CNNs is that ResNet has skip connections that allows it to skip layers. Fig. 1 describes the architecture of InceptionVasGloMyotides, where different layers and blocks are shown. Different blocks are opened in Figs. 2 and 3.

Data *normalization* is done, because data sources are different and different bioinformatic tests present the results in different scale. Normalization is scaled between 0 and 1. For normalization we used normalization layer, because it uses mean and variance of individual features, and calculation of those is used only training set.

For a *normalization* layer we calculate mean and variance of each feature in training data while preprocessing the data, which will be used to normalize data in training and validating. Normalization is scaled between −1 and 1, making it 0 centered. Max pooling layer with pool size and strides of 10×1 reduces the patient timeline of 100 years to 10 years where each row represents max value from 10-day interval. Reducing is done, because we are aware of the sparseness of the data and can assume that the same test is not done very often. Then there are convolution blocks and Inception residual Blocks and convolution layer. After these there is a max pooling layer to get max values of the features and a flattening layer makes it vector format with length of 3930, and as output layer we used Dense layer. Last layer is the dropout layer where rate is 0.01.

Convolution block in Fig. 2 contains two convolution layers: 10×1 and 1×10. Both convolution layers have four output filters, strides of 1×1 and padding set to the same, which means that output size is the same as input size. As activation function, we used *Hyperbolic Tangent* (Tanh). Difference of convolution block and Inception residual block in Fig. 3 is that Inception residual block has skip connection that allows the skipping of two convolution layers. Comparing to conventional ResNet models, they use quite often global average pooling method and *rectified linear unit* (ReLU) as activation function. These were not suitable as we have so sparse data that most of the data points have 0 values, and the global average would be always affected by those. ReLU on the other hand would cause our negative values to get 0 value even though they might be as relevant as positive values and we would lose important information.

We used as optimizer *Adaptive moment estimation* (Adam) and as a loss function, we used *categorical cross entropy* (CCE) and as other metrics *accuracy* (ACC). We chose to keep Adam's hyperparameters in default values in learning rate, beta 1, beta 2 and epsilon.

2.2. XGBoost

Chen and Guestrin [18] introduced the tree boosting system called XGBoost which is a scalable and highly performative with sparse data. In addition to this, XGBoost does not consume resource as much as CNNs.

XGBoost does not support a single patient's data as matrix, we

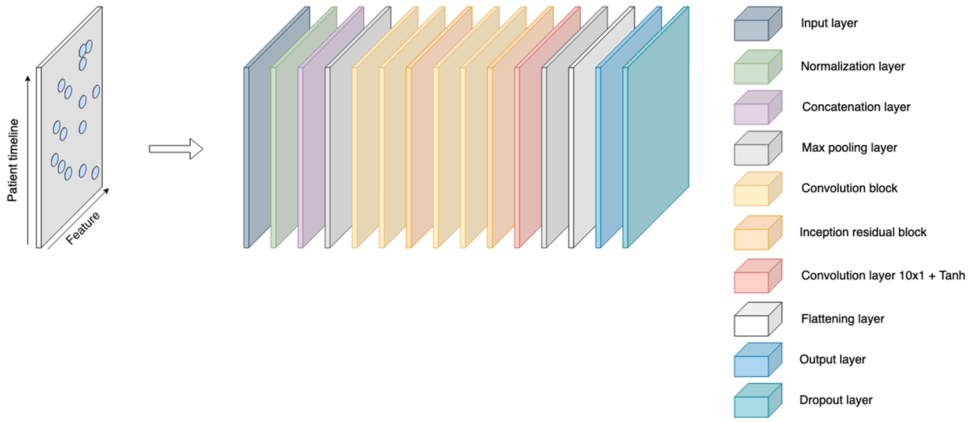

Fig. 1. Architecture of InceptionVasGloMyotides.

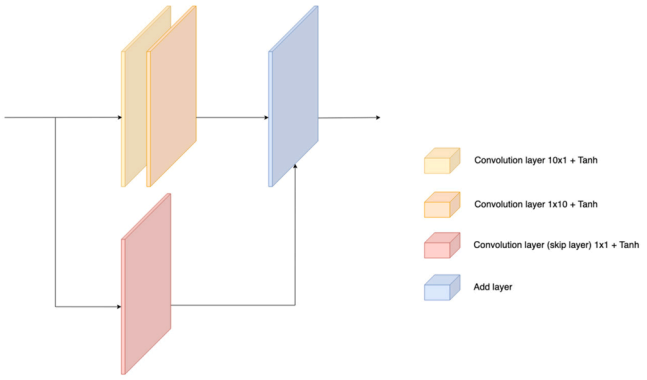

Fig. 2. Inception residual block of the InceptionVasGloMyotides.

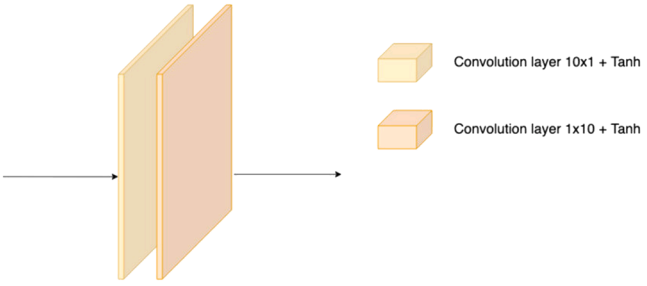

Fig. 3. Convolution block of the InceptionVasGloMyotides.

calculated minimum, maximum, mean and count for each feature. This results to vector of 15,720 features. We changed default hyperparameters of maximum depth (10), learning rate (0.05), L1 regularization (0.1), number of parallel trees constructed during each iteration (3) and learning task (multi class softprob). L1 requlariza tion is called *lasso regression* which adds penalty to the loss function. Hyperparameters were chosen by testing with grid search.

2.3. Research environment

In our setup, where we utilized two NVIDIA Tesla V100s graphics processing units, XGBoost's training required approximately one hour. InceptionVasGloMyotides model required approximately two hours for one epoch and maximum epochs we tested was 25 which took more than 2 days to finish [19].

2.4. Data

To secure enough data for ML approaches, we chose the above-mentioned, largest RD groups for further study, focusing on an imbalanced dataset of 114,897 patients, consisting of 100 000 randomly selected control objects, 2 919 vasculitiides (ICD-10: M30.0, M30.1, M31.3, M31.4, M31.7, D69.0, M36.4*D69.0 or N08.2*D69.0) patients, 942 myositides (ICD-10: M60.0, M60.1, M60.2, M60.8, M33.0, M33.1, M33.2, J99.1*M33.9, J99.1*M33.9, G72.41) patients and 11 036 glomerulonephritides (ICD-10: N0[0–9].*) patients. Then dataset is cleaned from null patients that do not have laboratory samples before structured RD diagnosis. This process can be seen in epidemical flowchart Fig. 4. Data for each patient and control include 1965 features of bioinformatics and 1965 features of numerical knowledge when bioinformatic value is out of range. In total, there were 3930 overall features from birth to current age or death. The most common features are *blood hemoglobin, blood leukocyte counts, red blood cell counts* and *hematocrit*. In the initial assessment, available data appeared sparse, but contained highly aberrant data on patient paths of studied patients versus controls.

2.5. Preprocessing

Data *transformation* follows the principles of tidy data, where columns are *variables*, rows are *observations* and *cells* contain a single value [20]. Our raw data format is long, which means that it needs to be pivoted to the wide format. At this point we needed to change the timeline from dates to number of days in the individual patients' life, e.g., date one is date of birth and as hard code the maximum day of 36, 500, becoming an artificially produced death day, if the patient did not decease before that.

Data *masking* happens in two ways in our research. The first masking technique is *pseudonymization* of sensitive information of patients. Pseudonymization process begins with a unique *social security number* (SSN). New SSNs are generated for patients, which makes it possible to combine data sources. The second technique is *nulling*, which removes the data completely and it is used, e.g., for first and last name, because we do not need that information, or to hide values to make predictions realistic with forced unavailability of the eventual correct RD diagnosis. Hiding values in this context means that we are nulling all the values after the timepoint of the day of correct disease resulting in variable mask sizes between 30 and 4320 days before the diagnoses.

For InceptionVasGloMyotides data normalization was not performed during preprocessing. Only the mean and variance were calculated in this step. These values were used in the ResNet's normalization layer. However, XGBoost did include a normalization step due to the different data format. We *split* the data into a separate training, validation, and test sets for ResNet, and for XGBoost we had a separate training and test

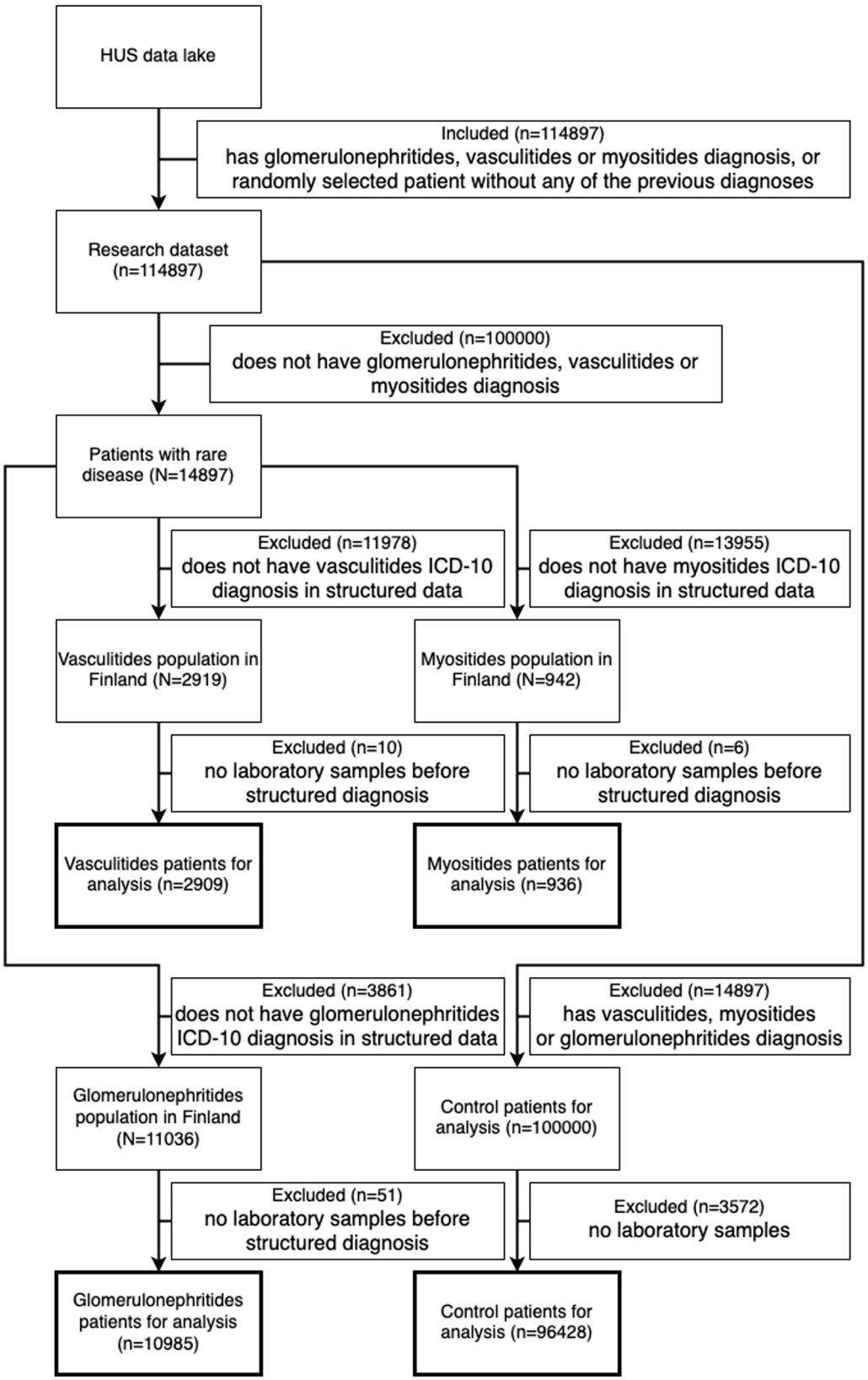

Fig. 4. Flowchart of epidemical design.

sets. In all cases, validation and test set sizes were 20 % of the full data set, and the rest of the data were for the training set.

2.6. Performance measures

Model performances were evaluated with multiple different metrics. Basic metrics *True Positive* (TP), *False Positive* (FP), *True Negative* (TN),

and *False Negative* (FN) were used in every formula. True positive ratio (TPR) describes the ratio of TPs over positives and true negative rate (TNR) describes the ratio of TNs over negatives, and False positive ratio (FPR) describes the ratio of FPs over negatives and False negative ratio (FNR) describe FNs over positives (1)-(4). *Area under the receiver operating characteristic curve* (AUC) value should vary between 50 and 100 %, with higher values implicating better performance. This value was received from the *Receiver Operating Characteristic* (ROC) curve which described the ratio of TPR and FPR. ACC (5) simply describes the ratio of correct classification over all classified objects.

$$TPR = \frac{TP}{TP + FN} \times 100\ \% \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \times 100\ \% \tag{2}$$

$$TNR = \frac{TN}{TN + FP} \times 100\ \% \tag{3}$$

$$FNR = \frac{FN}{FN + TP} \times 100\ \% \tag{4}$$

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \times 100\ \% \tag{5}$$

Positive predictive value (PPV) described the ratio of patients truly diagnosed as positive to all those who had a positive algorithm result (6). Negative predictive value (NPV) described the ratio of those truly negative to those who had a negative algorithm result (7). The formulas which considered disease prevalence for PPV and NPV were designed as pPPV (8) and pNPV (9). Considering population prevalence gives a more exact estimate of the likelihood of finding the correct diagnosis [19]. Threshold describes where the PPVs and NPVs were reached, and it informed us what should be used as the baseline of prediction certainty to classify patient with RD.

$$PPV = \frac{TP}{TP + FP} \times 100\ \% \tag{6}$$

$$NPV = \frac{TN}{TN + FN} \times 100\ \% \tag{7}$$

$$pPPV = \frac{TPR \times prevalence}{TPR \times prevalence + (1 - TNR) \times (1 - prevalence)} \times 100\ \% \tag{8}$$

$$pNPV = \frac{TNR \times (1 - prevalence)}{(1 - TPR) \times prevalence + TNR \times (1 - prevalence)} \times 100\ \% \tag{9}$$

3. Results

With InceptionVasGloMyotides model the highest sensitivities for binary classification (i.e., patient had at least one of the studied diseases), vasculitides and glomerulonephritides were reached in 30-days masking shown in Table 1. Myosisitides obtained their highest TPR when 4320-day masking was applied. Binary TNR achieved its highest value in 4320-day masking, as did vasculitides and glomerulonephritides. However, the highest TNR for myositides was reached with 2160-day masking. In addition to these masking sizes, there were tests with 0-, 120-, 360-, 720-, 1440-, 2880-days masking.

Table 2 lists various PPVs and corresponding NPVs versus specific thresholds. Notably, when prevalence was not considered, the highest PPVs in most cases were in 4230-days masking, where binary classification had 99.7 %, vasculitides had 90.0 % and glomerulonephritides had 98.1 %. Myositides did not reach PPV above 90 %. The highest NPVs did not reach above 90 % in the most cases, but for myositides it was 93.2 % in 2160-day masking, and vasculitides reached a decent 85.0 % in the 4320-days masking. Thresholds were lowest in the 30-days masking, excluding binary classification in the 4320-days masking. When the prevalence was considered for 2160 days, masking had the highest scores, where binary classification pPPV was 6 %, vasculitides was 0.2 % and myositides was 0.3 %, and for glomerulonephritides the highest pPPV was of 0.5 % in the 30-days masking [19].

Table 1
TPR, TNR, AUC, and ACC of the InceptionVasGloMyotides (%).

	TPR	TNR	AUC	ACC
30 days masking				
Binary	92.5	63.7	88.7	82
Vasculitides	82.5	89.7	72.6	88
Myositides	85.1	78.9	77.3	79
Glomerulonephritides	89.4	84.3	84.7	86
2160 days masking				
Binary	89.8	80.2	93.0	86
Vasculitides	79.3	95.3	76.3	93
Myositides	77.1	89.3	78.5	88
Glomerulonephritides	86.9	92.8	87.1	90
4320 days masking				
Binary	91.5	81.0	94.5	88
Vasculitides	78.9	96.4	76.3	94
Myositides	87.1	88.3	82.2	88
Glomerulonephritides	88.0	94.0	87.1	91

Table 3 shows similar results for XGBoost as InceptionVasGloMyotides reached. TPR had the highest probabilities for binary classification, and for mysitides and glomerulonephritides in the 30-days masking. The highest value for vasculitides was in the 2160-days masking. The highest TNR probabilities for binary classification, and for myositides and glomerulonephritides, were in the 30-days masking. Vasculitides had the highest value in the 2160-days masking.

Table 4 shows that binary classification and individual disease classifications reached PPVs above 96 % in all high score cases. Vasculitides had 96.7 % and myositides 97.5 % in 4320-days masking. Glomerulonephritides had 97.1 % in 2160 days masking and binary classification with 99.8 % was in 30 days masking. Majority of NPVs were under 90 % except for myositides, reaching 96.5 % in 30-days masking. Vasculitis had reached a high value of 89.0 % in 2160-days masking. All the highest pPPVs and pNPVs calculated were in the 30-days masking. The highest pPPV for binary classification was 8.8 %, vasculitides had pPPV of 0.6 %, myositides reached 1 % and glomerulonephritides had 0.5 %, and NPVS all over 99.98 % [19].

4. Related works

Compared to other published DDSS and ML applications in single RDs, our binary approach was approximately comparable or better, potentially due to analysis of higher numbers of affected. Jia et al. [15] developed the RDAD system, an ML system to support phenotype-based RD diagnostics. They showed PPV values reaching 99 % with up to 95 % TPR. If for comparison our PPVs were calculated by using a non-prevalence-corrected version, we reached roughly equal PPV results to RDAD's *phenotype based rare diseases similarity* (PICS) model for example when using the InceptionVasGloMyotides's 4320 days masking model in glomerulonephritides (98.1 %). At the same time, our model reached a higher TPR (88% vs. 62 %). In addition, in the 30 days masking model, our approach reached roughly similarly high PPV (99.6 %) and TPR (92.5 %) values [15]. Compared with other CNN models in clearly more common diseases with similar data construction, the reported AUC scores average between 70 and 75 % in *Chronic Obstructive Pulmonary Disease* (COPD) and *Congestive Heart Failure* (CHF). AUC in our InceptionVasGloMyotides model averaged around 80 %, reaching 92 % with binary classification [21]. Thus, when scanning for rare events, complex diseases may need lower numbers of known patients than if more common diseases were scanned.

Compared to Yoo et al. [22] conjunctival melanoma detection, which is very different task than ours, but it is having the same objective of

Table 2
Highest PPV and pPPV, and NPV and pNPV in the same threshold received with InceptionVasGloMyotides (%).

	PPV	NPV	Threshold	pPPV	pNPV	Threshold
30 days masking						
Binary	99.6	37.9	83.7	4	99.985	99
Vasculitides	76.7	77.9	82.2	0.1	99.99	86
Myositides	50.0	92.5	89.6	0.2	99.998	90
Glomerulonephritides	93.3	48.4	91.6	0.5	99.995	90
2160 days masking						
Binary	99.6	39.1	95.5	6	99.987	99
Vasculitides	78.8	84.6	92.6	0.2	99.99	97
Myositides	80.0	93.2	98.9	0.3	99.998	94
Glomerulonephritides	97.3	48.6	99.3	0.3	99.995	98
4320 days masking						
Binary	99.7	39.8	77.4	6	99.986	99
Vasculitides	90.0	85.0	99.5	0.1	99.99	72
Myositides	50.0	92.5	96.4	0.05	99.998	77
Glomerulonephritides	98.1	51.2	98.9	0.4	99.995	88

Table 3
TPR, TNR, AUC, and ACC of the XGBoost (%).

	TPR	TNR	AUC	ACC
30 days masking				
Binary	91.3	89.5	96.4	91
Vasculitides	76.7	98.4	80.8	95
Myositides	66.7	99.7	75.7	98
Glomerulonephritides	91.6	91.0	89.6	91
2160 days masking				
Binary	91.1	89.0	96.4	90
Vasculitides	79.5	98.6	82.1	96
Myositides	65.6	99.4	69.9	97
Glomerulonephritides	91.2	90.6	90.0	91
4320 days masking				
Binary	90.2	87.9	96.3	89
Vasculitides	75.2	98.2	79.9	94
Myositides	61.7	99.5	71.2	97
Glomerulonephritides	91.0	89.8	89.5	90

improving early identification. They accomplished ACC of 81.0 % with MobileNetV2 multiclass classification which would compare to our individual disease ACCs varying from 79 % to 94 % and averaging in 82.6 % across all the masking's with InceptionVasGloMyotides deep learning model. XGBoost ACCs varies between 90 % and 98 %, and averages in 94.3 %. AUC in the other hand falls behind in both our models compared to the MobileNetV2. Binary classification also falls behind with both models in all results that can be compared against MobileNetV2.

5. Discussion & conclusions

In this study, we demonstrated that our binary classification model outperformed all disease group specific classifications. Binary classification strategy in related inflammatory diseases could thus potentially be used in expedited AI-assisted diagnostic consultations. Classification of individual diseases also reached competitive levels, as XGBoost reached PPVs over 90 %. Also, InceptionVasGloMyotides reached in most of the PPVs values above 90 %, but not as consistently as XGBoost. NPV results were similar and above binary classification NPVs regardless of whether XGBoost or InceptionVasGloMyotides was used. In conclusion, simultaneous scanning of complex, related inflammatory diseases for expedited assessment by devoted specialists seems potentially feasible.

A narrower masking (30 days) in general resulted in better TPR values than other retrospective masking strategies. TPR of glomerulonephritides was higher than myositides and vasculitides suggesting that glomerulonephritides' disease progression may be more disease specific and in the future easier to pinpoint by DDSS. The fact that all masking strategies, regardless of their lengths, reached surprisingly high sensitivities suggests that the natural progression of all these diseases was slow and clinically insidious, while there may be differences between the disease groups in when they come clinically apparent by the used model.

Interestingly, when we compared the state-of-the-art XGBoost to InceptionVasGloMyotides, the latter model performed better with more extensive data masking, while XGBoost was better with less masking. This suggests that InceptionVasGloMyotides could in future become more effective in earlier discovery of an ongoing disease process. However, any differences in results were judged to be rather marginal, while InceptionVasGloMyotides model appeared very competitive against XGBoost. The biggest known difference is the required training time: XGBoost does not require much computational power.

A weakness in our work was to choose optimization of PPVs (over

Table 4
Highest PPV and pPPV, and NPV and pNPV in the same threshold received with XGBoost (%).

	PPV	NPV	Threshold	pPPV	pNPV	Threshold
30 days masking						
Binary	99.8	39.4	98.4	8.8	99.989	99
Vasculitides	92.3	86.7	96.6	0.6	99.99	97
Myositides	96.2	96.5	93.0	1	99.999	93
Glomerulonephritides	94.7	44.9	99.6	0.5	99.996	99
2160 days masking						
Binary	99.7	39.3	98.8	4	99.987	99
Vasculitides	96.3	89.0	96.6	0.2	99.99	98
Myositides	94.7	95.3	96.9	1	99.999	92
Glomerulonephritides	97.1	47.3	99.3	0.2	99.995	99
4320 days masking						
Binary	99.7	39.5	98.3	6	99.987	99
Vasculitides	96.7	85.2	99.0	0.1	99.99	98
Myositides	97.5	95.1	93.8	1	99.999	92
Glomerulonephritides	96.2	46.6	99.8	0.2	99.995	99

NPVs), instead of selecting the best possible means for both variables. Such optimization by lowering PPVs would result in increasing NPVs, which here reached less satisfactory results. In designing DDSS during prospective studies, one will always have to balance TPR vs. TNR, i.e., in effect to decide ethically, which one is more desirable and causes less net inefficiency, false alerts, unnecessary clinical procedures while optimizing the net decrease in disease-specific human suffering. In addition to this we have data limitations, where we are depending on structured diagnosis data that have some patients who have received the RD diagnosis before the source systems are taken to use. From these situations we have learned that patient journals can have in some cases indications that RD disease has been diagnosed earlier than it is in structured data. Also, there are patients outside from HUS area, which means that they came with doctor's referral and do not have sufficient data for prediction. These issues have been tackled with masking of the data because both cases have first RD diagnosis is very early and these patients do not have many laboratory results before that diagnosis date, therefore it cleans the most of these cases out of the data.

Needed future studies include developing, configuring, and honing these models to reach performance improvements. The used 2-step classification (binary and disease specific) seems enticing to introduce into more widespread use, as here binary classification seemed very accurate, and could be employed as the first level to filter RD patients from other patients. In the second level, one could classify the most probable RD if other criteria will be met or give information of the likelihood of various RDs. Also, one possible research line in the future could be few-shot learning, which have been proven effective with ResNet-style network in the rare fungus disease diagnosis [23].

Statement of ethical approval

The research was approved by the *Institutional Review Board* (IRB), and Finnish Social and Health Data Permit Authority Findata has approved secondary use of patient data. The latest permit number is THL/4465/14.06.00/2022.

Declaration of Competing Interest

The authors declare that they have no financial or personal relationships with other people or organizations that can inappropriately influence our study.

Acknowledgments

We would like to thank Business Finland for funding this eCare for Me research project, Helsinki University Hospital for coordination of this project and its collaboration ecosystem CleverHealth Netowork, and Tietoevry Ltd for partnership and providing the technological knowledge and capabilities.

References

[1] A. So, D. Hooshyar, K.W. Park, H.S. Lim, Early diagnosis of dementia from clinical data by machine learning techniques, Applied Sciences 7 (2017) 651, https://doi.org/10.3390/app7070651 (Switzerland).

[2] F. Shen, S. Liu, Y. Wang, L. Wang, N. Afzal, H. Liu, Leveraging collaborative filtering to accelerate rare disease diagnosis, in: Proceedings of the AMIA Annual Symposium Proceedings, 2017, pp. 1554–1563.

[3] C. Molster, D. Urwin, L. Di Pietro, M. Fookes, D. Petrie, S. Van Der Laan, H. Dawkins, Survey of healthcare experiences of Australian adults living with rare diseases, Orphanet. J. Rare. Dis. 11 (2016) 30, https://doi.org/10.1186/s13023-016-0409-z.

[4] Y. Zurynski, M. Deverell, T. Dalkeith, S. Johnson, J. Christodoulou, H. Leonard, E. J. Elliott, Australian children living with rare diseases: experiences of diagnosis and perceived consequences of diagnostic delays, Orphanet. J. Rare. Dis. 12 (2017) 68, https://doi.org/10.1186/s13023-017-0622-4.

[5] C.J. Hendriksz, Rare disease impact report: insights from patients and the medical community, 2022 (2013). https://www.researchgate.net/publication/236982217_Rare_Disease_Impact_Report_Insights_from_patients_and_the_medical_community (accessed April 5, 2022).

[6] S.L. Sawyer, T. Hartley, D.A. Dyment, C.L. Beaulieu, J. Schwartzentruber, A. Smith, H.M. Bedford, G. Bernard, F.P. Bernier, B. Brais, D.E. Bulman, J. Warman Chardon, D. Chitayat, J. Deladoëy, B.A. Fernandez, P. Frosk, M.T. Geraghty, B. Gerull, W. Gibson, R.M. Gow, G.E. Graham, J.S. Green, E. Heon, G. Horvath, A.M. Innes, N. Jabado, R.H. Kim, R.K. Koenekoop, A. Khan, O.J. Lehmann, R. Mendoza-Londono, J.L. Michaud, S.M. Nikkel, L.S. Penney, C. Polychronakos, J. Richer, G. A. Rouleau, M.E. Samuels, V.M. Siu, O. Suchowersky, M.A. Tarnopolsky, G. Yoon, F.R. Zahir, J. Majewski, K.M. Boycott, Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care, Clin. Genet. 89 (2016) 275–284, https://doi.org/10.1111/cge.12654.

[7] A.C. Wu, P. McMahon, C. Lu, Ending the diagnostic odyssey - Is whole-genome sequencing the answer? JAMA Pediatr 174 (2020) 821–822, https://doi.org/10.1001/jamapediatrics.2020.1522.

[8] H. Chinoy, R.G. Cooper, Myositis, Oxford University Press, 2018. https://oxfordmedicine-com.libproxy.tuni.fi/view/10.1093/med/9780198754121.001.0001/med-9780198754121.

[9] Orphanet, Inclusion body myositis, 2022 (2022). https://www.orpha.net/consor/cgi-bin/OC_Exp.php?Lng=GB&Expert=611 (accessed April 11, 2022).

[10] Orphanet, Polymyositis, 2022 (2022). https://www.orpha.net/consor/cgi-bin/OC_Exp.php?Lng=GB&Expert=732 (accessed April 11, 2022).

[11] Orphanet, Dermatomyositis, 2022 (2022). https://www.orpha.net/consor/cgi-bin/OC_Exp.php?lng=EN&Expert=221 (accessed April 11, 2022).

[12] P. Davey, D. Sprigings, Diagnosis and Treatment in Internal Medicine, Oxford University Press, 2018. https://oxfordmedicine-com.libproxy.tuni.fi/view/10.1093/med/9780199568741.001.0001/med-9780199568741.

[13] Orphanet, Vasculitis, 2022 (2022). https://www.orpha.net/consor/cgi-bin/OC_Exp.php?Expert=52759&lng=EN (accessed April 11, 2022).

[14] K.T. Woo, C.M. Chan, Y.M. Chin, H.L. Choong, H.K. Tan, M. Foo, V. Anantharaman, G.S.L. Lee, G.S.C. Chiang, P.H. Tan, C.H. Lim, C.C. Tan, E. Lee, H.B. Tan, S. Fook-Chong, Y.K. Lau, K.S. Wong, Global evolutionary trend of the prevalence of primary glomerulonephritis over the past three decades, Nephron Clin. Pract. 116 (2010) 337–346, https://doi.org/10.1159/000319594.

[15] J. Jia, R. Wang, Z. An, Y. Guo, X. Ni, T. Shi, RDAD: a machine learning system to support phenotype-based rare disease diagnosis, Front. Genet.. 9 (2018) 587. 10.3389/fgene.2018.00587.

[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[17] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.A. Muller, Deep learning for time series classification: a review, Data Min Knowl Discov 33 (2019) 917–963, https://doi.org/10.1007/s10618-019-00619-1.

[18] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.

[19] R. Ryyppö, Residual Neural Network in the Identification of Rare Diseases, Tampere University, 2021.

[20] H. Wickham, Tidy data, J. Stat. Softw. 59 (2014) 1–23.

[21] Y. Cheng, F. Wang, P. Zhang, J. Hu, Risk prediction with electronic health records: a deep learning approach, in: Proceedings of the 16th SIAM International Conference on Data Mining 2016, Society for Industrial and Applied Mathematics, 2016, pp. 432–440, https://doi.org/10.1137/1.9781611974348.49.

[22] T.K. Yoo, J.Y. Choi, H.K. Kim, I.H. Ryu, J.K. Kim, Adopting low-shot deep learning for the detection of conjunctival melanoma using ocular surface images, Comput. Methods Programs Biomed. 205 (2021), https://doi.org/10.1016/j.cmpb.2021.106086.

[23] M. Gao, H. Jiang, L. Zhu, Z. Jiang, M. Geng, Q. Ren, Y. Lu, Discriminative ensemble meta-learning with co-regularization for rare fundus diseases diagnosis, Med. Image Anal. 89 (2023), https://doi.org/10.1016/j.media.2023.102884.