



Research Paper

Artificial intelligence aided serum protein electrophoresis analysis of Finnish patient samples: Retrospective validation

Tapio Lahtiharju^{a,*}, Lassi Paavolainen^b, Janne Suvisaari^a, Pasi Nokelainen^a, Emmi Rotgers^c, Mikko Anttonen^a, Outi Itkonen^a

^a Department of Clinical Chemistry, HUS Diagnostic Centre, Helsinki University Hospital and University of Helsinki, P.O. Box 720, FI-00029 HUS, Finland

^b Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, P.O. Box 20, FI-00014, Finland

^c Fimlab Laboratories Oy Ltd, P.O. Box 66, FI-33013, Finland

ARTICLE INFO

Keywords:

Artificial intelligence
Blood protein electrophoresis
Deep learning
Multiple myeloma
Paraproteinemia
Paraproteins

ABSTRACT

Background and aims: Serum protein electrophoresis interpretation requires a substantial amount of manual work. In 2020, Chabrun *et al.* created a machine learning method called SPECTR for the task. We aimed to validate and test the SPECTR method against our results of more precise immunofixation electrophoresis.

Materials and methods: We gathered 34 625 patients and their first serum protein electrophoresis sample in Helsinki University Hospital. We trained three neural network models: (1) a fractionation model to fractionate electropherograms; (2) a classification model to classify samples to normal, ambiguous, and abnormal (*i.e.* containing paraprotein); (3) an integration model to predict concentration and location of paraproteins.

Results: The fractionation model demonstrated an error rate of ≤ 0.33 g/L in 95 % samples. The classification model achieved an area under the curve of 97 % in receiver operating characteristic analysis. The integration model demonstrated a coefficient of determination (R^2) of 0.991 and a root-mean-square error of 1.37 g/L in linear regression.

Conclusion: The neural network models proved to be suitable for partial automation in serum protein electrophoresis reporting, *i.e.* classification of normal electropherograms. Furthermore, the models can accurately suggest the location and concentration of paraproteins.

1. Introduction

Multiple myeloma is a relatively common malignancy of the bone marrow. It has a global age-standard rate of incidence 1.8 per 100 000 people [1]. Serum protein electrophoresis (SPE) is a cornerstone laboratory test for screening, diagnosis, and follow-up of myeloma, Waldenström's macroglobulinaemia, monoclonal gammopathy of undetermined significance, and other monoclonal gammopathies [2]. SPE involves many manual steps, which is a major drawback due to the time consumed by laboratory professionals and the potential for human error.

In SPE, serum proteins are separated into albumin, alpha1, alpha2, beta1, beta2, and gamma fractions in an electrophoresis matrix. The main factors affecting separation depend on the proteins' characteristics, *e.g.* the charge and size, or the electrophoretic conditions like the

strength of electric field, ion concentration and pH. For detection, the protein fractions are either stained in gel matrix or detected photometrically in capillary electrophoresis (CE) [3]. Traditionally, the resulting gels or electropherograms are manually interpreted by clinical laboratory specialists.

Machine learning methods have been developed to automate tasks in clinical laboratories [4]. Studies of neural network interpreting of SPEs date back to at least 1992 [5]. They were first used only for classifying samples to normal and abnormal [5–8]. In 2020, Chabrun *et al.* [9] reported of a machine learning method called SPECTR to interpret capillary electropherograms. Based on two datasets from French laboratories SPECTR was able to predict fractions, paraproteins, beta-gamma bridging, and restriction of heterogeneity with an accuracy comparable to an expert interpretation. In all SPECTR models an expert's

Abbreviations: CE, Capillary electrophoresis; CM, Classification Model; FM, Fractionation Model; HUS, Helsinki University Hospital; IFE, Immunofixation electrophoresis; PIM, Paraprotein Integration Model; SPE, Serum protein electrophoresis; TG, Training group; VG, Validation group.

* Corresponding author.

E-mail addresses: tapio.lahtiharju@hus.fi (T. Lahtiharju), lassi.paavolainen@helsinki.fi (L. Paavolainen), janne.suvisaari@hus.fi (J. Suvisaari), pasi.nokelainen@hus.fi (P. Nokelainen), emmi.rotgers@fimlab.fi (E. Rotgers), mikko.anttonen@hus.fi (M. Anttonen), outi.itkonen@hus.fi (O. Itkonen).

<https://doi.org/10.1016/j.cca.2024.120086>

Received 10 October 2024; Received in revised form 5 December 2024; Accepted 8 December 2024

Available online 9 December 2024

0009-8981/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

interpretation of SPE was used as the ground truth. Sensitivity and specificity of SPECTR outperformed earlier neural network models trained on capillary electropherograms [5–9].

In addition, two previous studies described neural network models trained with densitograms of gel images. While one detects only the presence of paraproteins [10], the other model also detects presence of acute phase reaction, hypoproteinaemia, nephrotic syndrome, and polyclonal gammopathy [11]. In addition to neural networks, mathematical formulae and decision trees have been employed to classify samples [12].

We aimed to validate and test SPECTR method with patient samples from Helsinki University Hospital (HUS) and to further develop the method by using only the first samples to emulate screening. Additionally, for paraprotein detection, the gold standard method, immunofixation electrophoresis (IFE), was used as a ground truth.

2. Materials and methods

2.1. Patient samples

In this retrospective study, we employed 34 625 patients and samples collected from February 2014 to January 2021 (Fig. 1). For each patient, only the first sample in HUS was included. The samples had been drawn into serum gel-barrier tubes from major manufacturers (e.g. 5/3.5 ml tube from Becton Dickinson, Franklin Lakes, New Jersey, USA; Ref 368498).

The samples were divided into a validation group (VG) and a training group (TG). First, we randomly separated 8 656 (25 %) of the samples to the VG. These results were not used to train any model. TG-PI and VG-PI includes only the samples having paraprotein concentrations estimated

by integration (PI) (Fig. 1).

2.2. Biochemical Assays

In our laboratory we report the concentrations of protein fractions and possible paraproteins. The electropherograms are interpreted using Phoresis software (Sebia, Lisses, France). The sample is forwarded to IFE, if there is a suspicion of new paraprotein, a paraprotein detected prior has disappeared, or the test is directly ordered by the clinician. If a paraprotein is detected in the sample, it is quantified using the perpendicular drop method [13]. However, the whole peak is not integrated; instead, the polyclonal background is estimated, and the width of the integration is reduced to equal the estimated concentration of the paraprotein alone. In training the models, we used the smoothed curves of electropherograms. The smoothing was conducted by the Phoresis software with a factor of 2, which is the default factory setting.

The samples were analysed with Capillarys 3 Tera or Capillarys 2 capillary electrophoresis instruments using Capi 3 Proteine 6 assay (Sebia). In addition to the electropherogram data, the results included details of fractionation, paraprotein integration, protein concentration, and the report of the IFE.

All samples were analysed for serum total protein by biuret reaction with the following analysers: Roche Hitachi Modular PE (F. Hoffmann-La Roche, Basel, Switzerland) by Total Protein method during 2014–2016, with Abbott Architect c16000 (Abbott Laboratories, Abbott Park, Illinois, United States) by Total Protein method during 2016–2019, and with Siemens Atellica Solutions (Siemens Healthineers, Erlangen, Germany), Atellica CH930 Total Protein II during 2019–2021. The serum total protein concentration was used to calculate the concentrations of the protein fractions and paraproteins.

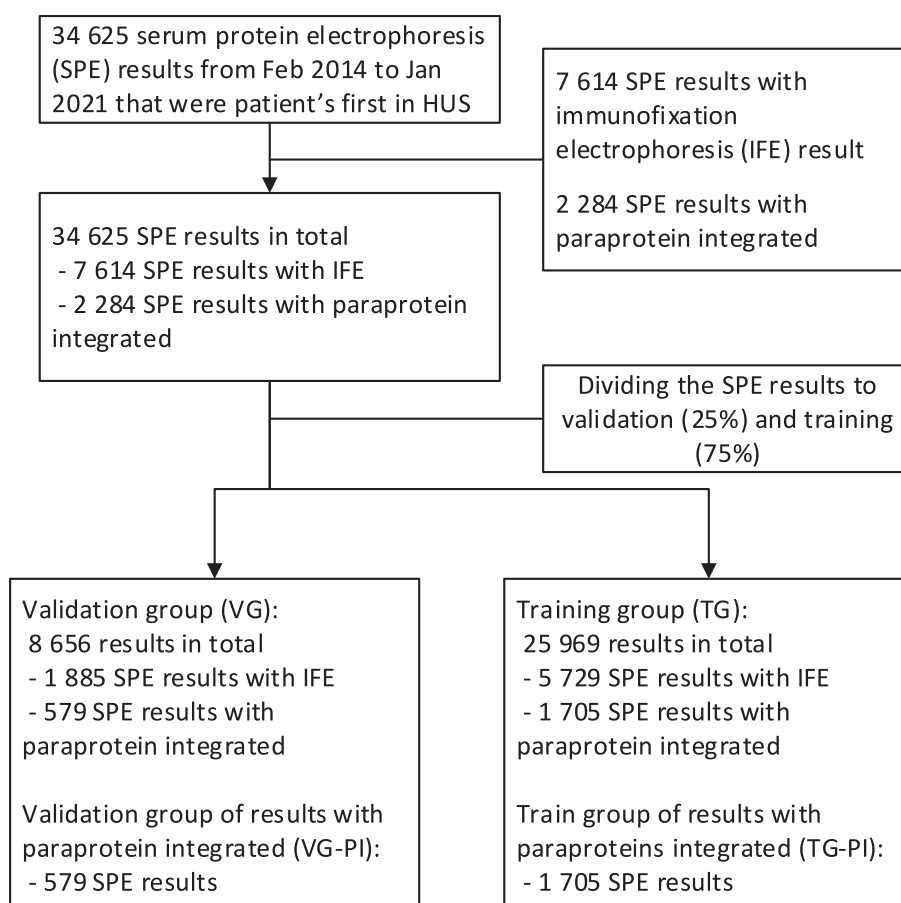


Fig. 1. The flowchart of the selection of the samples and group selections.

CE findings of 7 614 (22 %) samples had been confirmed by IFE using Hydrigel 4 or 9 IF reagents (Sebia) on Hydrasys 2 Scan instruments (Sebia) (Fig. 1). We searched from the IFE reports computationally with RegEx rules if the patient had paraprotein in the sample. Confirmation remained uncertain in 117 (1.5 %) reports, and these were considered normal *i.e.* not containing a paraprotein. The reliability of the computational classification was assured by selecting 100 random reports and manually checking that the reports were correctly classified. All the classifications were correct, thus, with the binomial exact confidence intervals we can assume that 96.4–100 % (95 % confidence interval (CI)) were correctly classified.

All samples were analysed in HUS Diagnostic Centre, the accredited (SFS-EN ISO 15189:2007 during 2014–2016 and SFS-EN ISO 15189:2013 during 2016–2021) clinical chemistry laboratory of HUS.

2.3. The neural network models

The three neural network models are all based on SPECTR, a machine learning method developed by Chabrun et al. The Supplement chapter 1. presents exact details of the models' architecture, while this section outlines the significant changes when compared to SPECTR.

2.3.1. Fractionation model

Fractionation Model (FM) classifies each point in the CE input curve as one of six fractions: albumin, alpha-1, alpha-2, beta-1, beta-2, and gamma. It is a derivative of the fractioning model of SPECTR except that FM assigns each point in the curve to one of the aforementioned fractions. The ground truth data for training FM was based on the fractionation marks set in Sebia's Phoresis program, and for training we used TG. Each mark in the curve indicates a start of a new zone. In total, five fractionation marks separated the six different zones fractioned in the ground truth data. These marks were initially generated automatically by the Phoresis program and then reviewed by a laboratory specialist for any necessary manual adjustments. A chi-squared test was used to determine whether there were significant differences in the number of errors between samples with and without paraprotein in the ground truth.

2.3.2. Classification model

Classification Model (CM) predicts whether a CE curve is normal or abnormal. It is based on the classification model of SPECTR except that CM was trained only to detect paraproteins. CM was trained with TG, where we used as the ground truth the result of IFE. If a sample did not have an IFE result, it was considered normal.

The outcome of CM is two values within the range of 0.0–1.0: one for the prediction of normality and one for the prediction of abnormality. The outcome is directly proportional to the confidence of the model, with higher values indicating higher confidence. The prediction of normality is equal to one minus the prediction of abnormality. Therefore, a sum of CM outcomes is always 1.0. The result is considered normal or abnormal if the confidence of the prediction is over the selected threshold. Otherwise, the result is considered as ambiguous, meaning that the model is not confident in either normal or abnormal prediction. When analysing receiver operating characteristic (ROC) area under curve (AUC) and calculating sensitivity and specificity with different thresholds, ambiguous results were considered as abnormal.

In addition to the analyses described above, we created two other models for the classification. The architecture was the same for these models, but they were trained by the following subset of samples: 1) samples not including IFE ordered by clinician, and 2) samples for which IFE was done. In 1) the ground truth was considered abnormal if IFE was done.

To ensure the absence of bias in the division of samples, cross-validation was conducted. The random division of samples into training and validation groups was repeated ten times, with a separate training and validation performed for each iteration. The mean and

confidence interval were calculated for the results assuming normal distribution, and these values were then compared to those obtained with CM.

2.3.3. Paraprotein integration model

Paraprotein Integration Model (PIM) predicts for each point in the electropherogram whether it belongs to a paraprotein or not. It is based on the peak detection model of SPECTR with no major differences. We trained PIM using curves of TG-PI, where paraprotein(s) had been integrated by laboratory specialists. The model was not trained with normal samples as the paraprotein integration model is only used to locate paraproteins in samples classified as abnormal. A chi-squared test was used to compare integration errors of paraproteins in the gamma fraction with those in the other fractions, based on the number of samples divided according to the error points.

2.4. Software

Patient sample data was processed in HUS Acamedic, which is a secure, scalable, virtual, and audited operating environment for safe processing of sensitive data [14]. All model training and evaluation were implemented in Jupyter Notebook environment using Python 3.8.10. Main Python packages used in the project were TensorFlow 2.11.0 for model compiling, training, and prediction, NumPy 1.23.5 for data processing, scikit-learn 1.2.1 for outcome evaluation, pandas 1.5.3 for data management, and matplotlib 3.6.3 and Seaborn 0.12.2 for visualization. The code for model architectures and training setup was modified from that of SPECTR [9,15].

2.5. Ethics

The study was approved by the Medical Research Committee of HUS (§26/2022) and was conducted in accordance with the ethical principles of the Declaration of Helsinki. In accordance with the ethical standards governing research, no other approval was required on account of the secondary nature of the research. In reporting, the study adhered to STARD guidelines [16].

3. Results

3.1. The patients

The total number of samples and patients was 34,625, with 52 % women and a median age of 69 years (Table 1). Of the patients with a paraprotein integrated in CE, 46 % were women, with a median age of 73 years. The ground truth by IFE was normal in 88 % of samples in both TG and VG, and in 7 % of TG-PI and 8 % of VG-PI. There were no significant differences in age, gender distribution, or location of paraproteins between TG, VG, TG-PI, and VG-PI.

3.2. Fractionation model

FM predicted the location of fractions (albumin – gamma). The performance of the model was evaluated with respect to predicted concentrations of the protein fractions and the location of the fractions. The errors were ≤ 0.33 g/L in 95 % of the samples in VG (Fig. 2). This includes the errors in all fractions. With normal ground truth in classification, 95 % of the samples had an error ≤ 0.29 g/L and, with abnormal ground truth, 95 % of the samples had an error ≤ 1.3 g/L (Fig. 2). We also compared the proportion of samples with error points (each point equals 1/300 of the electropherogram) (Table S1). The error was ≤ 2 points in 94 % in all samples, in 95 % of samples with normal ground truth in classification, and in 86 % of samples with abnormal ground truth. The errors of concentrations were similar across the fractions (Figure S1). FM performed better for samples with normal than for those with abnormal classification ground truth in terms of number

Table 1

Patient and sample characteristics of the training groups (TG, TG-PI) and validation groups (VG, VG-PI). Location of paraproteins presents the proportion of all samples with an integrated paraprotein in the given location. Because a paraprotein can expand into multiple fractions, the total sum for each row exceeds 100 %.

Group	n	Age		Women	Ground truth		Location of paraproteins					
		Median [IQR]			Normal	Abnormal	Albumin	Alpha-1	Alpha-2	Beta-1	Beta-2	Gamma
TG	25 969	69 [57–77]		52 %	88 %	12 %	0.1 %	0 %	0.4 %	3 %	11 %	96 %
VG	8 656	68 [56–77]		52 %	88 %	12 %	0 %	0 %	0.2 %	3 %	13 %	94 %
TG-PI	1 705	73 [65–81]		46 %	6.8 %	93 %	0.1 %	0 %	0.4 %	3 %	11 %	96 %
VG-PI	579	73 [65–80]		45 %	7.9 %	92 %	0 %	0 %	0.2 %	3 %	13 %	94 %

Abbreviations. TG – training group of all samples and patients, VG – validation group of all samples and patients, TG-PI – training group of samples and patients with paraprotein integrated, VG-PI – validation group of samples and patients with paraprotein integrated, IQR – inter-quartile range.

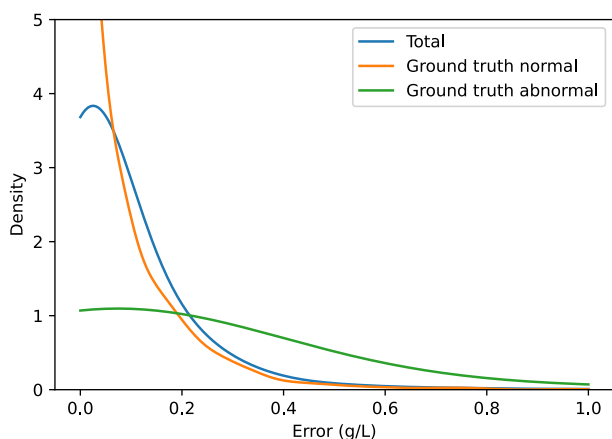


Fig. 2. The concentration error density plots for Fractionation Model in total VG, in VG with ground truth normal, and in VG with ground truth abnormal.

of error points and concentration (p -value $< 1e-06$) (Fig. 2, Table S1). In the case of VG samples with paraprotein integrated, the median error was 0.1 g/L, with an inter-quartile range (IQR) of 0.0–0.1 g/L. This was observed when the paraproteins did not coincide the fraction changes. In the event of coincidence, the error was 2.0 g/L (IQR 1.3–6.9 g/L). This difference is significant (p -value $< 1e-06$). There were 27 samples (0.3 % from VG) that FM predicted ineligible fractions: either an incorrect count or incorrect order. These were excluded from the analysis of FM.

3.3. Classification model

CM predicted whether the samples were normal, ambiguous, or abnormal (*i.e.* presence of paraprotein in the sample). In ROC analysis of normal versus ambiguous or abnormal cases the AUC was 97 % (95 % CI 96–97 %) (Figure S2), and with a threshold of 0.95, a sensitivity of 95.2 % and a specificity of 81.2 % were achieved (Table 2). With a threshold of 0.8, the precision was 98.3 % for normal and 88.5 % for abnormal outcome. Tightening the threshold to 0.975 improves the precision to 99.5 % and 92.8 %, respectively, but simultaneously increases the proportion of ambiguous results, *i.e.* the results that were neither above the normal nor abnormal threshold, from 8.2 % to 33 % (Fig. 3, Table 2). With a threshold of 0.95, precision of CM was 99.2 % for normal and

Table 2

The results of Classification Model with different thresholds. The proportion of model outcomes in the validation group shows the percentage of samples classified as normal, ambiguous, and abnormal. The precision describes proportion of correctly classified samples when compared to the ground truth. Sensitivity and specificity are calculated from the normal threshold, *i.e.* ambiguous is considered abnormal.

Threshold	Proportion of model outcomes			Precision		Sensitivity	Specificity
	Normal	Ambiguous	Abnormal	Normal	Abnormal		
0.80	82.9 %	8.2 %	8.8 %	98.3 %	88.5 %	87.9 %	92.5 %
0.90	78.3 %	14.3 %	7.4 %	98.7 %	90.1 %	91.5 %	87.7 %
0.95	72.1 %	22.3 %	5.6 %	99.2 %	91.7 %	95.2 %	81.2 %
0.975	64.1 %	33.0 %	2.9 %	99.5 %	92.8 %	97.5 %	72.4 %

91.7 % for abnormal results. Using this threshold, 49 samples (0.8 %) with abnormal ground truth were incorrectly classified as normal. A closer look at the IFE reports of these samples revealed that the paraprotein concentration was < 1 g/L in 35 (71 %) samples and < 5 g/L in 46 (94 %) samples (Table S2). We also tested training the models with samples without clinician ordered IFEs or with only the samples for which IFEs were performed, but these did not improve the outcomes (Tables S3, S4 and S5). Finally, in the cross-validation, the results of CM were found to be well within the confidence intervals, and thus no significant bias was detected (Table S6).

3.4. Paraprotein integration model

PIM predicted the location and concentration of paraproteins. In 95 % of the samples in VG the error was ≤ 1.6 g/L (Fig. 4A). Comparison of paraprotein concentrations between the ground truth and predicted by PIM by linear regression shows coefficient of determination (R^2) 0.991 and root-mean-square error (RMSE) 1.37 g/L (Fig. 4B). The intercept was -0.14 (95 % CI 0.002 to -0.29) g/L, and the slope 1.005 (0.997 to 1.012). In addition, we calculated the errors as the number of points. The error was ≤ 2 points in 75 % of all samples, in 82 % of the samples with a paraprotein in the gamma fraction, and in 67 % of the samples with a paraprotein in another fraction than gamma (Table S7). The model performed better when the paraprotein was in the gamma fraction (p -value 0.01).

4. Discussion

We have confirmed that the principle in SPECTR developed by Chabrun *et al.* [9] in fractionating, classifying normal and integrating paraprotein SPE electropherograms is effective and reliable. We used patient samples from HUS, and the gold standard IFE as the ground truth. To improve precision of our models to detect paraproteins, we defined an ambiguous group with different thresholds to account for samples that are not definitively normal or abnormal.

Our fractionation model FM performed better than that of SPECTR. While SPECTR had a standard deviation of 0.91 g/L and 1.17 g/L for gamma fraction in internal and external test sets [9], respectively, our model had deviation less than 0.33 g/L in 95 % of the samples. Furthermore, our findings demonstrate that when a paraprotein coincides with a fraction change, the error rate of fraction concentrations

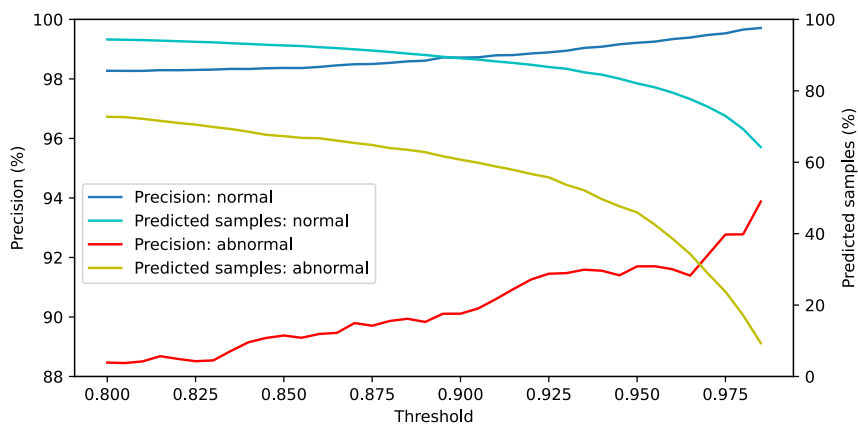


Fig. 3. Performance of Classification Model at different thresholds. In the left axis, there is precision of the classification normal ground truth as normal (blue) and abnormal ground truth as abnormal (red). In the right axis, there is proportion of normal samples predicted as normal (teal) and proportion of abnormal samples predicted as abnormal (yellow).

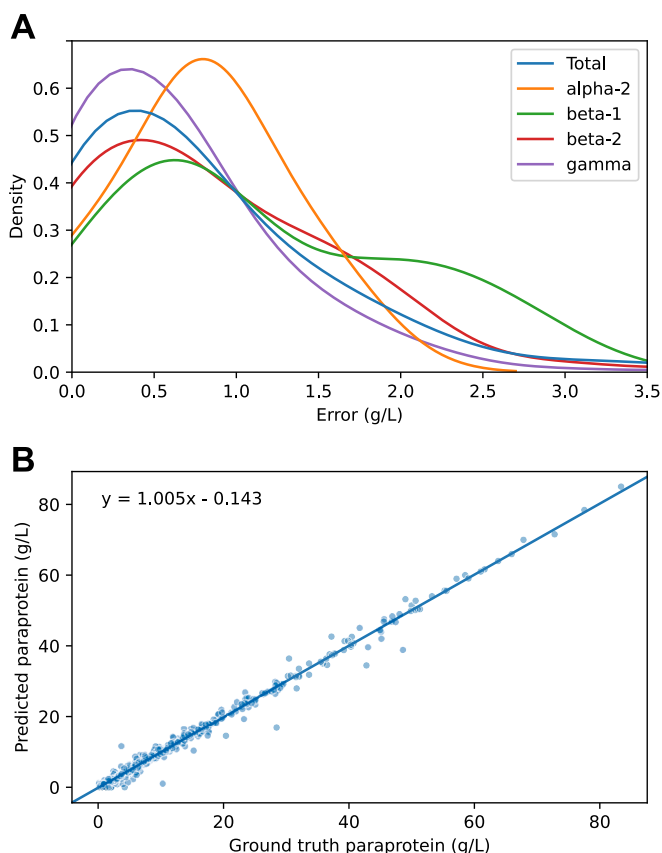


Fig. 4. Paraprotein Integration Model's performance with validation samples (VG-PI). A) Density plot of error concentrations in total and in different protein fractions. B) Linear regression analysis of the prediction and the ground truth concentrations.

is markedly elevated. Based on our clinical experience, in addition to posing a challenge for our model, these situations are also challenging for human interpreters.

Our classification model CM outperforms previous machine learning models trained on gel samples [10,11], but is slightly inferior to the original SPECTR [9]: SPECTR had sensitivity of 97.8 % and 95.5 %, and specificity of 96.6 % and 96.2 % in internal and external test sets, respectively [17]. Our model had sensitivity of 91.5 % and specificity of 87.7 % with the same threshold of 0.90. Similarly, our model had AUC

97 %, while SPECTR had AUC > 99 % and 99 % with internal and external test sets, respectively. The fact that our ground truth was based on IFE rather than laboratory specialist's opinion of the SPE is likely to explain the difference. We chose IFE as the ground truth as it is the gold standard for the detection of paraproteins. Using IFE as the ground truth likely decreases the observed specificity since samples with unspecific peaks caused by for example high CRP or other interferences are not classified as abnormal by the ground truth method. Additionally, our decision to include only the first sample from each patient may have contributed to this discrepancy, as the first samples may be more difficult to classify. Furthermore, our dataset was smaller with higher average patient age as compared to that of Chabrun *et al.* [9].

Jonsson *et al.* [12] developed a decision algorithm to classify samples according to the results of selected mathematical formulae. The evaluation was based on gel electrophoresis interpretation ($n = 711$) as the ground truth. There were 95 samples with monoclonal immunoglobulins and nine samples with monoclonal free light chains. The sensitivity to detect was 98.9 % with the monoclonal immunoglobulins and 56 % with the monoclonal free light chains. The specificity was 99.5 % and the detection limit in gamma region 1 g/L. However, compared to our study the sample size was small, and the ground truth defined differently.

The sensitivity of CE has been estimated to be 98 % and the specificity 91 % when compared to immunoelectrophoresis or immunofixation [18]. With a threshold of 0.975, our CM has similar sensitivity, but specificity is 72 %. This may partly be explained by differing proportions of samples with paraprotein concentrations ≤ 1 g/L, but also demonstrates that there is still room for further improvement of AI models.

The results from our additional models indicate that CM has superior precision to CM-EO, which has been trained using the interpreter's suspicion of paraprotein as the ground truth, rather than the IFE (Table S5). Therefore, the findings suggest that the model may outperform the average interpreter.

Paraprotein integration with our PIM model performed equally well to the original SPECTR. When comparing model predictions to ground truth both studies had the same coefficient of determination of 0.99. For our PIM RMSE was 1.37 g/L, and for SPECTR it was 1.23 g/L with internal test set and 1.13 g/L with external test set [9]. Our study provides further evidence that the methodology used in SPECTR is valid and robust, even if compared to the gold standard IFE, HUS patient samples, and our reporting style. We also demonstrate that the samples with paraprotein in other regions than gamma are more prone to errors. This was expected as they are also more challenging to detect and integrate by laboratory specialists.

Our study has limitations that need to be considered. First, we chose the first sample of each patient to represent new cases. However, some

patients may have had a plasma cell disorder diagnosed at an earlier date not included in our data, and the first sample in our registry was in fact a follow-up sample. Most of these cases are due to previous mergers of laboratories in our region. Secondly, initial fractionating of electropherograms is currently made automatically by Sebia's Phoresis program that searches the local minima. If the initial fractionating is incorrect, the laboratory specialist should correct it manually. In practice, this may remain undone and as always, other human errors are possible as well. However, these are likely to apply to only a small proportion of the samples and are thus unlikely to have a significant impact on our overall results.

Although our sample size was above 34 000 samples, it is not enough to train AI models with rare or very rare conditions. For instance, paraproteins occur infrequently in the alpha fractions and therefore errors in this fraction are more common. Jacobs *et al.* [19] have demonstrated that the detection limit of SPE is around 1 g/L. This is in line with our finding that samples misclassified by CM often had ≤ 1 g/L paraprotein. A small proportion of errors can be tolerated because clinicians are aware of the limitations of SPE in detecting some paraproteins. When paraproteins are strongly suspected, clinicians often order IFE directly. Taken together, CM can be used in clinical practice to automatically classify samples. In our hospital, 88 % of new SPE samples are interpreted as normal. In the future, CM could be used to reduce the manual work of laboratory specialists by automatically identifying normal results for release to the laboratory information system.

5. Conclusions

This study provides further evidence that machine learning models could partly replace manual interpretation of SPE results. Our models can reliably fractionate and classify normal samples and could therefore be exploited to select results for automatic reporting to electronic health records. In addition, the models can facilitate fractionating and integrating paraproteins in ambiguous or abnormal cases. Besides speeding up the process, AI models are also likely to reduce human errors in interpretation and help to direct the expert's attention to challenging or pathological cases. However, further work is needed to improve specificity of the present models especially in rare paraproteinaemias. New models to predict re-appearance of the original paraprotein during clinical follow-up, or the incorporation of additional patient data to the predictive models, e.g. serum free light chains, could further improve the accuracy and reduce the need for confirmation by the gold standard IFE. Lastly, the final judgment by the clinical laboratory specialist remains.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used Microsoft Word proofing tools and DeepL Write in order to improve language. Generative AI was not used. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRedit authorship contribution statement

Tapio Lahtiharju: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Investigation, Formal analysis, Data curation. **Lassi Paavola:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation. **Janne Suvisaari:** Writing – review & editing, Validation, Conceptualization. **Pasi Nokelainen:** Writing – review & editing, Validation, Conceptualization. **Emmi Rotgers:** Writing – review & editing, Validation, Conceptualization. **Mikko Anttonen:** Writing – review & editing, Validation, Supervision, Conceptualization. **Outi Itkonen:** Writing – review & editing, Validation, Supervision, Project administration, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank Riku Turkki for his assistance in initiating and planning the project, HUS Data Lake and Data Production for providing us with the patient data, and HUS Academic for providing us with the secure operating environment. This work was supported by funding from HUS Diagnostic Centre during 2019–2024; the Research Council of Finland [grant number 340273] for L.P.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cca.2024.120086>.

Data availability

The data that support the findings of this study are available upon request from the corresponding author, T.L. The data are not publicly available due to the secondary nature of the study.

References

- [1] J. Huang, S.C. Chan, V. Lok, L. Zhang, D.E. Lucero-Priso, W. Xu, Z.-J. Zheng, E. Elcarte, M. Withers, M.C.S. Wong, The epidemiological landscape of multiple myeloma: a global cancer registry estimate of disease burden, risk factors, and temporal trends, *Lancet Haematol.* 9 (2022) e670–e677, [https://doi.org/10.1016/S2352-3026\(22\)00165-X](https://doi.org/10.1016/S2352-3026(22)00165-X).
- [2] The International Myeloma Working Group, Criteria for the classification of monoclonal gammopathies, multiple myeloma and related disorders: a report of the International Myeloma Working Group, *Br. J. Haematol.* 121 (2003) 749–757, <https://doi.org/10.1046/j.1365-2141.2003.04355.x>.
- [3] M.A. Jenkins, M.D. Guerin, Quantification of serum proteins using capillary electrophoresis, *Ann. Clin. Biochem.* 32 (Pt 5) (1995) 493–497, <https://doi.org/10.1177/000456329503200510>.
- [4] S.R. Master, T.C. Badrick, A. Bietenbeck, S. Haymond, Machine Learning in Laboratory Medicine: Recommendations of the IFCC Working Group, *Clin. Chem.* 69 (2023) 690–698, <https://doi.org/10.1093/clinchem/hvad055>.
- [5] M.A. Kratzer, B. Ivandic, A. Fateh-Moghadam, Neuronal network analysis of serum electrophoresis, *J. Clin. Pathol.* 45 (1992) 612–615, <https://doi.org/10.1136/jcp.45.7.612>.
- [6] A. Ognibene, R. Motta, A. Caldini, A. Terreni, E.D. Dea, M. Fabris, G. Messeri, Artificial neural network-based algorithm for the evaluation of serum protein capillary electrophoresis, *Clin. Chem. Lab. Med.* 42 (2004) 1451–1452, <https://doi.org/10.1515/CCLM.2004.271>.
- [7] S. Altinier, L. Sarti, M. Varagnolo, M. Zaninotto, M. Maggini, M. Plebani, An expert system for the classification of serum protein electrophoresis patterns, *Clin. Chem. Lab. Med.* 46 (2008) 1458–1463, <https://doi.org/10.1515/CCLM.2008.284>.
- [8] A. Ognibene, M.S. Graziani, A. Caldini, A. Terreni, G. Righetti, M.C. Varagnolo, A. Campanella, M. Martelli, R. Mancini, P. Rizzotti, M. Plebani, M. Mori, G. Gaspari, R. Motta, G. Galli, M. Fabris, G. Messeri, Computer-assisted detection of monoclonal components: results from the multicenter study for the evaluation of CASPER (Computer Assisted Serum Protein Electrophoresis Recognizer) algorithm, *Clin. Chem. Lab. Med.* 46 (2008) 1183–1188, <https://doi.org/10.1515/CCLM.2008.221>.
- [9] F. Chabrun, X. Dieu, M. Ferre, O. Gaillard, A. Mery, J.M. Chao de la Barca, A. Taisne, G. Urbanski, P. Reynier, D. Mirebeau-Prunier, Achieving Expert-Level Interpretation of Serum Protein Electrophoresis through Deep Learning Driven by Human Reasoning, *Clin. Chem.* 67 (2021) 1406–1414, <https://doi.org/10.1093/clinchem/hvab133>.
- [10] R. Chen, D.L. Jaye, J.D. Roback, M.A. Sherman, G.H. Smith, Lightweight, open source, easy-use algorithm and web service for paraprotein screening using spatial frequency domain analysis of electrophoresis studies, *J. Pathol. Inform.* 13 (2022) 100128, <https://doi.org/10.1016/j.jpi.2022.100128>.
- [11] N. Lee, S. Jeong, K. Jeon, W. Song, M.-J. Park, Development and validation of a deep learning-based protein electrophoresis classification algorithm, *PLOS ONE* 17 (2022) e0273284.
- [12] M. Jonsson, J. Carlson, J.-O. Jeppsson, P. Simonsson, Computer-supported Detection of M-Components and Evaluation of Immunoglobulins after Capillary Electrophoresis, *Clin. Chem.* 47 (2001) 110–117, <https://doi.org/10.1093/clinchem/47.1.110>.

- [13] C. Schild, B. Wermuth, D. Trapp-Chiappini, F. Egger, J.-M. Nuoffer, Reliability of M protein quantification: comparison of two peak integration methods on Capillars 2, *Clin. Chem. Lab. Med.* 46 (2008) 876–877, <https://doi.org/10.1515/CCLM.2008.146>.
- [14] E. Turunen, HUS Academic. Virtual presentation., (2022). <https://sway.office.com/iiBUeFL0ZCF3ChY2> (accessed June 5, 2024).
- [15] F. Chabrun, X. Dieu, M. Ferre, O. Gaillard, A. Mery, J.M. Chao de la Barca, A. Taisne, G. Urbanski, P. Reynier, D. Mirebeau-Prunier, SPECTR: Serum Protein Electrophoresis Computer-Assisted Recognition AI, (2024). https://github.com/fchabrun/SPECTR_AI (accessed June 5, 2024).
- [16] J.F. Cohen, D.A. Korevaar, D.G. Altman, D.E. Bruns, C.A. Gatsonis, L. Hoof, L. Irwig, D. Levine, J.B. Reitsma, H.C.W. de Vet, P.M.M. Bossuyt, STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration, *BMJ Open* 6 (2016) e012799.
- [17] F. Chabrun, X. Dieu, P. Reynier, D. Mirebeau-Prunier, In Reply to Performance of Deep Learning in the Interpretation of Serum Protein Electrophoresis, *Clin. Chem.* 68 (2022) 1341–1343, <https://doi.org/10.1093/clinchem/hvac145>.
- [18] J.A. Katzmann, R. Clark, E. Wiegert, E. Sanders, R.P. Oda, R.A. Kyle, C. Namyst-Goldberg, J.P. Landers, Identification of monoclonal proteins in serum: a quantitative comparison of acetate, agarose gel, and capillary electrophoresis, *Electrophoresis* 18 (1997) 1775–1780, <https://doi.org/10.1002/elps.1150181011>.
- [19] J.F.M. Jacobs, K.A. Turner, M.S. Graziani, J.L. Frinack, M.W. Ettore, J.R. Tate, R. A. Booth, C.R. McCudden, D.F. Keren, J.C. Delgado, G. Zemtsovskaja, R. O. Fullinaw, A. Caldini, T. de Malmanche, K. Katakouzinou, M. Burke, G. Palladini, S. Altinier, M. Zaninotto, G. Righetti, M.T. Melki, S. Bell, M.A.V. Willrich, An international multi-center serum protein electrophoresis accuracy and M-protein isotyping study. Part II: limit of detection and follow-up of patients with small M-proteins, *Clin. Chem. Lab. Med.* CCLM 58 (2020) 547–559, <https://doi.org/10.1515/ccim-2019-1105>.