



Original Research Article

Lightweight, open source, easy-use algorithm and web service for paraprotein screening using spatial frequency domain analysis of electrophoresis studies

Robert Chen ^{*}, David L. Jaye, John D. Roback, Melanie A. Sherman, Geoffrey H. Smith

Department of Pathology and Laboratory Medicine, Emory University School of Medicine, Atlanta, GA, USA



ARTICLE INFO

Keywords:

Signal processing
Data processing
Electrophoresis
Laboratory methods and tools
Cancer - hematological/lymphoma
Myeloma

ABSTRACT

Introduction: Serum protein electrophoresis (SPEP) is commonly used to detect monoclonal paraproteins to meet laboratory diagnostic criteria for plasma cell neoplasms. We propose an automated screening method for paraprotein detection that uses minimal computational resources for training and deployment.

Methods: A model screening for paraproteins based on the presence of high-frequency components in the spatial frequency spectrum of the SPEP densitometry curve was calibrated on a set of 330 samples, and evaluated on representative ($n = 110$) and external ($n = 1,321$) test sets. The model takes as input a patient's serum densitometry curve and a standardized control curve and outputs a prediction of whether a paraprotein is present. We built an interactive web application allowing users to easily perform paraprotein screening given inputs for densitometry curves, as well as a macro-enabled spreadsheet for easy automated screening.

Results: When tuned to maximize likelihood ratio with minimum sensitivity 0.90, the model achieved AUC 0.90, sensitivity 0.90, positive-predictive value 0.64, specificity 0.55, and accuracy 0.72 in the representative test set. In the external test set, the model achieved AUC 0.90, sensitivity 0.97, positive-predictive value 0.42, specificity 0.29, and accuracy 0.52. A subset analysis showed sensitivities of 0.90, 0.96, and 1.0 in detecting low (0.1–0.5 g/dL), medium (0.5–3.0 g/dL), and high paraprotein levels (≥ 3.0 g/dL), respectively. We have released a web service allowing users to score their own SPEP data, and also released the algorithm and application programming interface in an open-source package allowing users to customize the model to their needs.

Conclusions: We developed a proof of concept for an automated method for paraprotein screening using only the characteristics of the SPEP curve. Future work should focus on testing the method with other laboratory data including immunofixation gels, as well as incorporation of outside data sources including clinical data.

Introduction

The diagnostic criteria for plasma cell neoplasms such as multiple myeloma includes the presence of a paraprotein which may be identified by serum protein electrophoresis (SPEP).^{1,2} Paraprotein screening from SPEP samples often involves the identification of a peak in the gamma region of the sample's densitometry curve. Furthermore, even with the rise of newer methods of paraprotein screening such as mass spectrometry and serum free light chain assay (sFLC), SPEP remains the most widely used and available screening test in medical laboratories; as of 2021 the College of American Pathologists (CAP) Guidelines continue to strongly recommend the use of SPEP in initial screening steps for paraproteins.³ While SPEP interpretation has traditionally been performed manually, the process lends itself to automation due to structural commonalities in SPEP signals

across different samples. Thus, an automated paraprotein detection method has the potential to streamline clinical workflows in the clinical laboratory.

Studies in the past have proposed computational methods to detect paraproteins from SPEP signals. Neural network models have been previously proposed for identifying paraproteins from raw SPEP patterns^{4–7}; however, there are limited implementations of such methods that allow for users to easily score SPEP curves with minimal resources. Chabrun et al⁷ developed an online web-based application that uses a neural network model to predict the existence of a paraprotein; however, the tool requires uploading SPEP signals in the form of database objects generated by Microsoft Database, requires a large training set of over 162 000 samples and requires additional post-processing of data generated by the electrophoresis machinery, all of which prohibit ease of use. Furthermore, a drawback of neural network methods compared to simpler methods is that they

Abbreviations: SPEP, serum protein electrophoresis; ECG, electrocardiogram; EEG, electroencephalogram; LIS, laboratory information system; EHC, Emory Healthcare.

* Corresponding author at: 1364 Clifton Road NE, Atlanta, GA 30322, USA.

E-mail address: rchen25@emory.edu (R. Chen).

<http://dx.doi.org/10.1016/j.jpi.2022.100128>

Received 23 March 2022/Received in revised form 18 July 2022/Accepted 18 July 2022

Available online 25 July 2022

2153-3539/© 2022 The Author(s). Published by Elsevier Inc. on behalf of Association for Pathology Informatics. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

require extensive training, often requiring significant data, time, and computational resources including graphical processing units (GPUs), making it difficult to perform both training and inference in resource-poor settings. Previously, an add-on feature called Neurosoft was integrated into implementations of Sebia Phoresis software.^{8,9} While useful, Neurosoft is not implemented by default in Sebia systems, making it difficult to adopt widely.

Frequency domain analysis generally refers to the decomposition of a signal (e.g., time series, image) into the frequency spectrum, which is subsequently filtered, typically for purposes of removing noise from data.¹⁰ While relatively lacking in the specific use case of paraprotein detection from SPEPs, usage of the frequency domain has been widely performed for the interpretation of continuous temporal longitudinal signals, including electrocardiograms (ECGs)¹¹ and electroencephalograms (EEGs).¹² For electrocardiograms, applications include the measurement of characteristics of ventricular conduction such as QRS complex duration.¹³ The Fourier transform has been used for facilitating extraction of morphological features in the frequency domain which have been used to aid identification of abnormal patterns such as cardiac arrhythmias.¹⁴ In EEG analysis, features can be extracted from the frequency domain and correlated with epileptiform events.¹⁵ As SPEG signals embody similarities to ECG and EEG signals with respect to the continuous signal structure, we posit that using frequency domain analysis may be appropriate for the stratification of SPEG signals, with the main difference being that SPEG signals are captured in a spatial domain (along an axis ranging from a positive anode to negative cathode) rather than a time domain.

We propose a method based on spatial frequency domain analysis as an alternative to neural network-based models that typically require intensive computational resources and large amounts of training data and time to train. Our method is designed to achieve high likelihood ratio and sensitivity for paraprotein screening purposes, and lightweight enough to train on a basic computing setup without extensive GPU resources, while only requiring about 330 samples to calibrate the model parameters. Further, the simplicity of our model allows for ease of adoption across a wide variety of laboratory settings. Additionally, we have publicly released a web service and Excel spreadsheet allowing for ease of access to our methodology.

Materials and methods

In the SPEG gel analysis workflow, we run patients' samples on the same gel as control samples (samples negative for paraproteins), which serve as a background signal. We focus the problem to the gamma region of the sample's densitometry curve, where paraprotein peaks commonly occur. Most acquisition devices (e.g., Sebia HYDRASYS 2 SCAN) compute region delimiters automatically, allowing for us to isolate the region of interest easily. The only remaining computational task is to screen for the paraprotein via spatial frequency domain analysis.

In our method, we first isolate the gamma region of the sample densitometry curve, and then subtract the background signal from the control densitometry curve. Next, we compute the spatial frequency spectrum using a Fourier transform, and then apply a numerical screening threshold for high frequency components, whereby the presence of high frequency components stratifies samples as likely to have a paraprotein. We calibrated the thresholding parameters of the model on a calibration set and evaluated the method on 2 separate test sets.

In our laboratory, samples are processed on a 15-lane SPEG gel. Fig. 1 depicts an example of 1 gel in which some lanes contain paraproteins in the gamma region of the gel. A data acquisition device scans the gel and converts each lane into a continuous signal. We employ the Sebia HYDRASYS 2 SCAN (Sebia SA, Lisses, Ile-de-France) device which generates a 300-point signal curve in a standard hexadecimal format for subsequent manual interpretation.²³

We conducted our analysis on a cohort of laboratory samples from patients who underwent bone marrow biopsy and SPEG analysis as part of an evaluation for multiple myeloma at Emory Healthcare. In the study, samples were selected and labeled manually for paraprotein status.

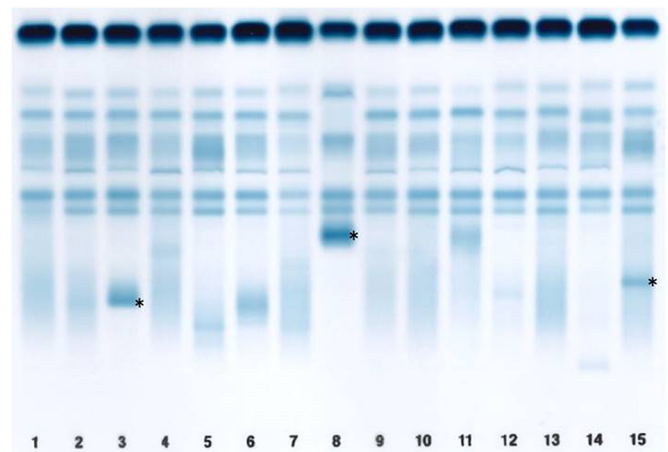


Fig. 1. A 15-lane SPEG gel from our clinical immunology laboratory. The anode is at the top and the cathode is at the bottom. We always run a normal serum control in lane 1 of every gel. Several patients have paraproteins in the gamma region on this gel, most prominently in lanes 8, 3, and 15 (asterisks).

Subsequently, a spatial frequency analysis model was calibrated on a validation set, evaluated on a representative test set of samples undergoing evaluation for multiple myeloma, and also evaluated on a separate, generalized external test set of samples that included all samples run through our laboratory in 1 month (Fig. 2).

Data

We curated primary data from a cohort (“Cohort 1”) of 3 relational databases associated with laboratory information systems (LIS) at Emory Healthcare (EHC): EHC Immunology Specialty LIS (Sebia Phoresis - PostgreSQL), EHC Clinical Pathology LIS (Cerner Millennium PathNet - Oracle RDBMS), EHC anatomic pathology LIS (Cerner CoPathPlus - Microsoft SQL Server). We extracted 7721 samples from a registry of patients known to have a plasma cell neoplasm and were undergoing bone marrow aspiration and hematological testing including serum protein electrophoresis. Among these, we included 6144 samples with an accession date between January 1, 2014 and August 20, 2019, all of which utilized the same laboratory platform for serum protein electrophoresis analysis. Additionally, we extracted a separate cohort (“Cohort 2”) of 1321 samples from the Emory Healthcare Immunology Specialty LIS representing all the cases run in February 2020. This cohort included patients with and without a known plasma cell neoplasm.

Determination of data labels

Among Cohort 1, we separated the samples into those with quantifiable paraprotein status (e.g., greater than 0 g/L) and those without quantifiable paraprotein status (e.g., 0 g/L, or undetectable). We randomly shuffled all cases in each group, and subsequently manually confirmed paraprotein status for 501 samples across both groups (243 samples with positive paraprotein status, and 258 samples with negative paraprotein status). Baseline characteristics of the patients from which samples were derived are included in Table 1.

Each sample was labeled based on manual analysis of the pathologist interpretation, which is recorded as free text in the laboratory information system. Samples with an accession date earlier than January 1, 2014 were filtered out, as our clinical laboratory changed laboratory information system providers at this time. Furthermore, samples with missing data for the densitometry curves were filtered out. After these filtering steps, the final Cohort 1 set included 440 samples (215 positive, 225 negative), which were randomly shuffled into a validation set ($n = 330$, 75% of Cohort 1 set) and a representative test set ($n = 110$, 25% of Cohort 1 set). For a separate supplemental analysis, we created a version of the validation set that did not include samples with paraproteins found outside of the

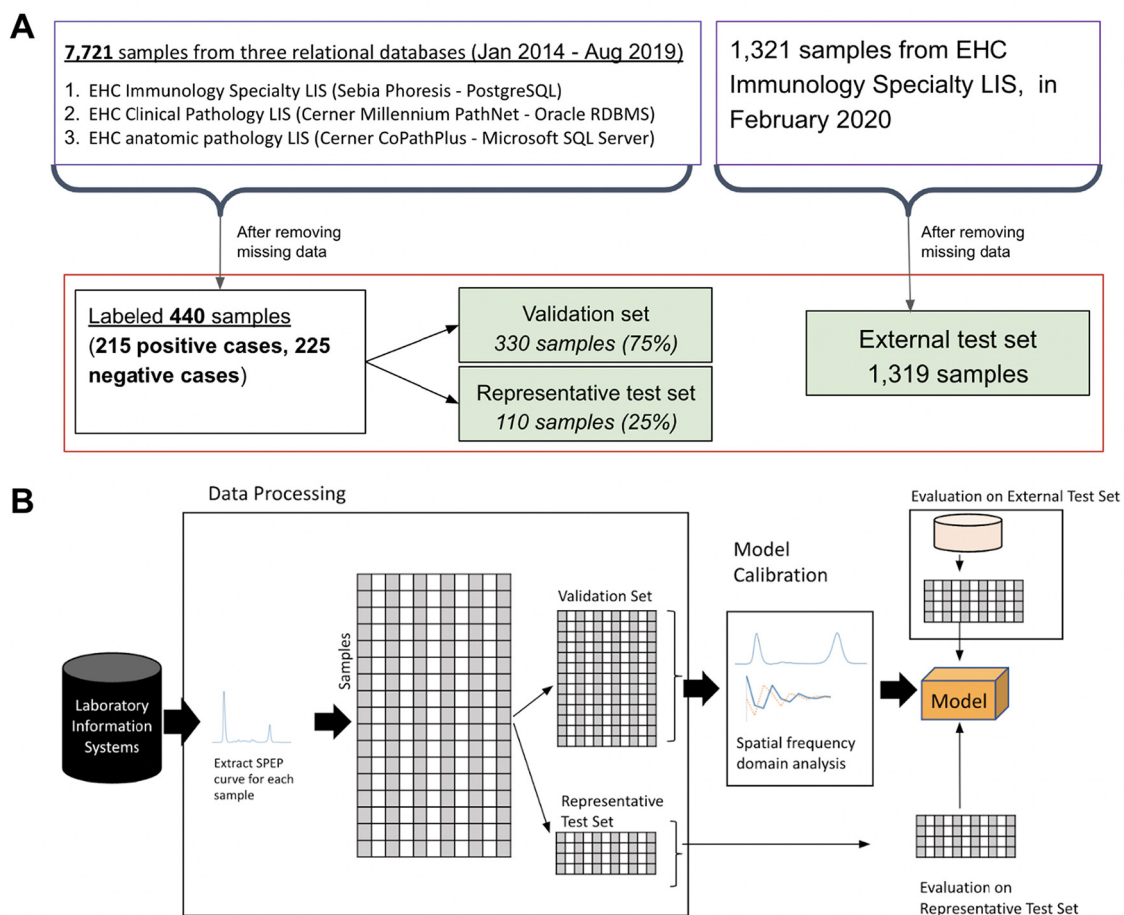


Fig. 2. (A) Framework for data processing, paraprotein screening model calibration, and evaluation. (B) Data sources and allocation of calibration and test sets. Calibration set samples were filtered out if they were collected before Jan 1, 2014, or had missing numerical data for the densitometry curve. External test set samples with missing data for quantifiable paraprotein concentration were filtered out.

gamma region (e.g., beta) to calibrate the model and assess the robustness compared to situations where paraproteins could only be in the gamma region (Supplemental Data, Table 1).

The samples from Cohort 2 were labeled according to quantified paraprotein status, where samples with quantified paraprotein concentrations less than 0.1 g/dL were considered to have negative paraprotein status while samples with quantified paraprotein concentrations of at least 0.1 g/dL were considered to have positive paraprotein status. The external test set was designed for the purpose of creating a large set of samples to test the model en masse. We used a 0.1 g/dL threshold for creating the external test set because we did not review the narrative interpretations for these samples and were unable to discern the paraprotein status as would have

been finalized by a pathologist. Since our methodology only considers the gamma region, it would be rare for any sample with a concentration below 0.1 g/dL to be an actual paraprotein. The entire set of samples from Cohort 2 was used as an external test set (Fig. 2B). Note that the external test set is considered a general set of samples whose purpose is to evaluate the portability of the algorithm to a disparate population from which the original model was trained.

Model

The model is a high-pass filter consisting of 2 successive steps. First, the spatial frequency spectrum of the raw SPEP curve is computed. Next,

Table 1

Baseline characteristics of the samples across the 2 paraprotein statuses (paraprotein positive and paraprotein negative). All values are reported as mean (standard deviation or percentage of population).

	Cohort 1 - validation set and representative test set	Paraprotein positive	Paraprotein negative
	<i>n</i> = 501	<i>n</i> = 243	<i>n</i> = 258
Age	63.2 (9.7)	62.8 (9.9)	63.5 (9.5)
Race, White [n (%)]	255 (57.2%)	116 (53.0%)	139 (61.2%)
Race, Black [n (%)]	160 (35.9%)	84 (38.4%)	76 (33.5%)
Gender, Male [(n (%)]	230 (51.6%)	116 (53.0%)	114 (50.2%)
Mean paraprotein concentration (g/L)	0.70 (1.39)	1.42 (1.70)	0.0 (0.0)
	Cohort 2 - external test set	Paraprotein positive	Paraprotein negative
	<i>n</i> = 1321	<i>n</i> = 458	<i>n</i> = 863
Mean paraprotein concentration (g/L)	0.34 (0.82)	0.97 (1.16)	0.0 (0.0)

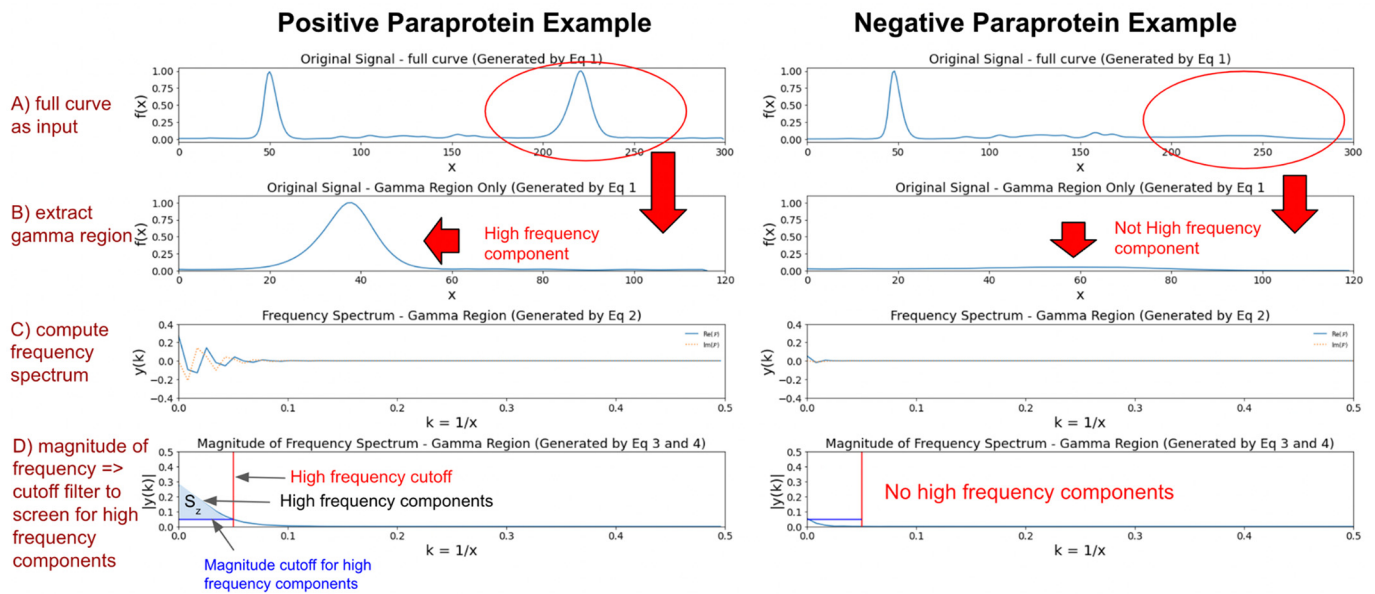


Fig. 3. Model characteristics. (A) The full curve with control curve subtracted (Equation 1). (B) The gamma region is extracted from the full curve (Equation 1, isolated to location of the gel where x is between $x_{\text{gamma,low}}$ and $x_{\text{gamma,high}}$). (C) The spatial frequency spectrum is computed (Equation 2). (D) Thresholds for spatial frequency and magnitude of the corresponding high frequency component is learned in the model calibration process (Equation 3, Equation 4). The learned model is subsequently scored on the representative and external test sets.

thresholds for a high-frequency cutoff and magnitude cutoff from the frequency spectrum are calibrated. We describe 2 models. The baseline “Serum Curve” model uses the raw SPEP curve as input (Equation 1a). The “Serum - Control” model uses the difference between the SPEP curve and the associated control curve as input (Equation 1). All steps afterward are the same for either model after the input is defined. For brevity, we describe the steps below using the “Serum-Control” model.

Problem formulation

For a particular sample, the gamma region for the SPEP densitometry curve was extracted (Fig. 3A: sample curve with control subtracted, generated by Equation 1; Fig. 3B: isolating the result of Equation 1 to just the gamma region where $\in [x_{\text{gamma,low}}, x_{\text{gamma,high}}]$) as follows:

Let $x \in [1, x_{\text{max}}]$ be the gel position of an SPEP curve. Let $\text{SPEP}_z(x)$ be the value of the patient SPEP curve for a particular sample z . Let $\text{CONTROL}_z(x)$ be the value of the control curve for that sample. In our samples, $x_{\text{max}} = 300$.

Let the input to the baseline “Serum Curve” model be

$$f_z(x) = \text{SPEP}_z(x). \tag{1a}$$

Let the input to the “Serum-Control” model be

$$f_z(x) = \text{SPEP}_z(x) - \text{CONTROL}_z(x). \tag{1}$$

Let $x \in [x_{\text{gamma,low}}, x_{\text{gamma,high}}]$ be the gel positions corresponding to the gamma fraction for the control sample associated with a specific sample. Note that the specific gamma region boundaries $x_{\text{gamma,low}}$ and $x_{\text{gamma,high}}$ are different for each sample and are computationally derived by the data acquisition platform (Sebia HYDRASYS 2 SCAN). Please note that every patient sample is run alongside a control, and every control sample contains a gamma fraction (i.e., it is not absent due to patient-specific characteristics like immunosuppression). While there may be variability across laboratories, in our laboratory we run 14 patient samples on a gel, all of which use the same control (i.e., lane 1 is a control and lanes 2–15 are patient samples and each sample is paired with the control on its gel; Fig. 1). Each different gel has a different control.

The fast Fourier transform^{24–26} for $f_z(x)$ (derived from Equation (1)) is derived with input $f_z(x)$, $x \in [x_{\text{gamma,low}}, x_{\text{gamma,high}}]$ (Fig. 3C):

$$y_z[k] = \sum_{x=x_{\text{gamma,low}}}^{N=x_{\text{gamma,high}}} \exp\left\{-2\pi j \frac{kx}{N}\right\} f_z[x], \tag{2}$$

Next, the magnitude of the real portion of the spatial frequency spectrum, $y_R = |y|$, is isolated (Fig. 3D), which has one value corresponding to each $k = 1/x, x \in [x_{\text{gamma,low}}, x_{\text{gamma,high}}]$. If the gel position x is in units of meters (m), then the spatial frequency position k is in units of $1/m$.

A high frequency component of $y_{z,R} = |y_z|$ is considered a component in the spatial frequency spectrum with cutoff frequency of q_{high} with a magnitude of at least m_{high} or

$$y_{z,R}[k] \text{ where } k \geq q_{\text{high}} \text{ and } y_{z,R}[k] \geq m_{\text{high}} \tag{3}$$

Let the individual score of a sample be S_z , defined as

$$S_z = \frac{\sum_{k=q_{\text{high}}}^{k_{\text{max}}} 1 \text{ if } |y_z[k]| > m_{\text{high}}, 0 \text{ otherwise}}{\sum_{k=q_{\text{high}}}^{k_{\text{max}}} 1}, \tag{4}$$

which corresponds to the proportion of elements in the high-frequency region of the frequency spectrum (spatial frequency above s) that have a magnitude above m_{high} (Fig. 3D).

Model training

All model parameter tuning is performed on the validation set using 10-fold cross validation described as follows. Samples are randomly evenly split into 10 subsets, and across 10 folds, the model is trained on 9 subsets and evaluated on the last subset. The final model is dictated by the parameters for the model corresponding to the fold with the highest area under the receiver operating curve (AUC).

The parameters for the model, q_{high} and m_{high} are tuned as follows, based on cutoffs that yield the highest area under the receiver operating curve (AUC) for classification of the sample. For given values of q_{high} and m_{high} ,

the AUC score is computed with values of $S_{z \in \{1, \dots, n_c\}}$ for all n_c samples in the validation set. Values of q_{high} ranging from $[1/x_{max}, 0.5]$ and m_{high} ranging from $(0, m_{max}]$ are used in the process, equally spaced across 1000 values within the ranges, where $1/x_{max}$ represents minimum possible spatial frequency corresponding to maximum gel position x_{max} and m_{max} represents the highest magnitude across all spatial frequency spectrums computed from all samples in the validation set. All combinations of q_{high} and m_{high} are assessed.

With the chosen q_{high} and m_{high} , a model threshold T for the scores is chosen to achieve a maximum likelihood ratio given a minimum sensitivity target, as computed among the validation data. To choose the threshold T , we construct a precision-recall curve using the scores S_z for all patient samples as predicted values and the ground-truth labels for paraprotein status as the true values. We pick the threshold that allows for maximum likelihood ratio given a minimum sensitivity target (e.g., given minimum sensitivity of 0.9). A sample z is classified as paraprotein positive if its score is greater or equal to the threshold, i.e., $S_z \geq T$ and is classified as paraprotein negative otherwise. Note that an alternate, separate threshold was computed in order to achieve a minimum sensitivity target, and is reported in (Supplemental Data, "Calibration of model to sensitivity").

Performance evaluation

Evaluation of the model was performed on both test sets as follows. Performance metrics were computed after calibrating the model to a maximum positive likelihood ratio (please note all subsequent mentions of "likelihood ratio" in the text refer to the positive likelihood ratio) given a minimum sensitivity of 0.90. If a model could not be chosen that has a viable operating point and still meets the criteria, the sensitivity cutoff was lowered until a model with a viable operating point could be obtained. That particular operating point served as the chosen model. We report an alternative evaluation calibrating the model to high sensitivity alone, in the Supplemental Data. While calibrating to high sensitivity is the primary objective for the goal of paraprotein screening, calibrating to the positive likelihood ratio allows for an optimal balance between maximizing detection and minimizing false positives. We believe that a minimum sensitivity of 0.90 is tolerable given that any samples determined to be negative by the model would go directly to the attending pathologist for interpretation while any samples determined to be positive would be pre-interpreted first before being read by a pathologist.

To determine the parameters of the model, q_{high} and m_{high} , we used 10-fold cross validation on the validation set and selected the values of q_{high} and m_{high} corresponding to the fold with the highest AUC value. Performance metrics were assessed on the representative test set and external test set using the selected model parameters. Metrics included area under the receiver operating curve (AUC), sensitivity, positive predictive value, specificity, F1-score and accuracy. For each test set, the AUC was computed using the range of thresholds $t \in [0, 1]$ applied to the model scored on the respective test set; sensitivity, positive predictive value, specificity, F1-score and accuracy were computed based on the threshold T that was calibrated on the validation set.

Separately, we conducted an analysis in which we manually inspected all densitometry curves for the positive paraprotein samples in the validation set, removed samples from the validation set if the paraprotein identified was outside of the gamma region, and assessed performance metrics with the subset of samples (reported in Supplemental Data, Table 1, Figs. 1, 2).

Web service

We developed a web service that allows for any client to input a raw serum densitometry curve and associated control densitometry curve in hexadecimal format and retrieve a prediction for whether the sample contains a paraprotein. The web service also provides an integer array representation of the sample curve and its associated control curve that can be used to plot the curves (Screenshot, Fig. 4).

A production version of our web service is provided at <https://trddx.emory.edu/spep/>. The code for the application programming interface

(API) is provided in the open source code repository (Supplemental Data, "Source code and usage guidance for model"). The web service is intended to be invoked by a computer program, such as our example Excel spreadsheet macro, but there is a test page that allows a user to interactively copy and paste data for control and sample curves in hexadecimal format and see the paraprotein prediction (1 = paraprotein detected, 0 = paraprotein not detected).

Excel spreadsheet for automated scoring

We developed a spreadsheet macro that invokes the web service, allowing a non-technical user to easily obtain a prediction for whether a sample contains a paraprotein, scored on our trained model. We provide an implementation in Microsoft Office Excel 365.

To leverage our tool, the user provides input in the form of raw sample and control gel densitometry curves, which can be queried from the PostgreSQL database used by the Sebia instrument (Fig. 5). The user may then score all samples and view the model prediction results directly in the spreadsheet (Fig. 6).

Technical implementation

All data processing was performed with core Python functions as well as the Pandas library.²⁷ For the model, the frequency spectrum $F(x)$ of the gamma region was computed using a Fourier transform from the `scipy.fft` Python library²⁶ and the validation and performance evaluation were performed using the `scikit-learn` Python library.²⁸ All steps of the process were performed on a 2017 MacBook Pro laptop computer with 2.3 GHz Dual-core Intel Core i5 processor, and integrated Graphics card (Intel Iris Plus Graphics 640, 1536 MB). The code for the model and web service, along with sample data, is provided publicly at <https://github.com/rchen25/spep-paraprotein-frequency-screen>.

Results

Final model parameters

We describe the results of the parameter tuning process, conducted via 10-fold cross validation. For the baseline "Serum Curve" model, the mean AUC across all folds of cross-validation was 0.915 (25th percentile: 0.912, 75th percentile: 0.919). The maximum AUC was 0.926. The final model parameter q_{high} corresponding to the maximum AUC was 0.030 (mean: 0.020, 25th percentile: 0.015, 75th percentile: 0.023), and the final model parameter m_{high} was 5.01E-4 (mean: 5.01E-4, 25th percentile: 5.01E-4, 75th percentile: 5.01E-4).

For the "Serum - Control" model, the mean AUC across all folds of cross-validation was 0.896 (25th percentile: 0.891, 75th percentile: 0.897). The maximum AUC was 0.912. The final model parameter q_{high} corresponding to maximum AUC was 0.010 (mean: 0.010, 25th percentile: 0.010, 75th percentile: 0.010), and the final model parameter m_{high} was 9.82E-4 (mean: 9.82E-4, 25th percentile: 9.82E-4, 75th percentile: 9.82E-4).

Performance evaluation on test sets

The baseline "Serum Curve" model achieved AUC of 0.93 on the representative test set and AUC of 0.88 on the external test set (Fig. 7). The "Serum-Control" model achieved AUC of 0.90 on the representative test set and AUC of 0.90 on the external test set (Fig. 7). While there is a drop in AUC between the representative test set and external test set in the baseline "Serum Curve" model, it is important to note that there is a lower proportion of samples with a paraprotein in the external test set. There is not a similar drop in AUC between representative test set and external test set in the "Serum-Control" model, which may indicate more robustness of this version of the model when using it in a test set from a population different than the one it was calibrated on.

seq	id	character varying(32)	nominativo character varying(30)	curva text
1	1	CTLNORMAL	021822 SERUM #1 WH	80000000000100010001000
2	2	CTLHYPERG	HYPERGAMMA	80020002000200020002000
3	3	003	7J	80010001000200020002000
4	4	003	1F	80050007000600050000400
5	5	001	7B	80010000000100010001000
6	6	003	1C	80010001000000000000000
7	7	003	7N	80010001000100010001000
8	8	003	2B	80010001000100010000000
9	9	002	7A	80010001000100010001000
10	10	003	8F	80020002000200020002000
11	11	003	4C	80000001000100020003000
12	12	003	0G	80010001000100010001000
13	13	003	6G	80010002000200020002000
14	14	002	1E	80010001000100020002000
15	15	002	5E	80020002000200010001000
16	16	CTLNORMAL	021822 SERUM #2 WH	80010001000100000000000
17	17	002	5E	80010001000100010000000
18	18	006	0G	80010002000200020002000
19	19	006	6A	80010000000000000000000
20	20	003	9D	80010000000000000000000
21	21	003	4F	80020001000100000000000
22	22	003	6C	80010000000000000000000

Fig. 5. Screen shot showing the retrieval of densitometric curve data for 1 full day's worth of samples run on Sebia HYDRASYS instruments in our clinical immunology laboratory. The samples are ordered by their position on the gels. Every gel consists of 15 lanes and we always run a normal serum control (CTLNORMAL) in the first lane. These data are retrieved from the PostgreSQL database associated with our Sebia Phoresis imaging system using standard query tools that Sebia provides with the system. Protected health information (patient names, "nominativo" column) are removed from the screenshot.

On the representative test set, the baseline "Serum Curve" model performed with sensitivity 0.90, positive-predictive value 0.70, specificity 0.66, F1 score 0.79, and accuracy 0.77. On the external test set, the baseline "Serum Curve" model performed with sensitivity 0.92, positive predictive-value 0.54, specificity 0.58, F1 score 0.68, and accuracy 0.70 (Table 2). On the representative test set, the "Serum-Control" model performed with sensitivity 0.90, positive-predictive value 0.64, specificity 0.55, F1 score 0.75, and accuracy 0.72. On the external test set, the "Serum-Control" model performed with sensitivity 0.97, positive-predictive value 0.42, specificity 0.29, F1 score 0.59, and accuracy 0.52 (Table 2).

In a separate analysis we have detailed the performance after manual removal of samples with a paraprotein outside of the gamma region (see Supplemental Data, Table 1, Figs. 1, 2).

Performance across strata of paraprotein concentrations

When calibrated to likelihood ratio, the external test set, the baseline "Serum Curve" model achieved sensitivity of 0.86 ($n = 216$) in the detection of low quantified paraprotein status (0.1–0.5 g/dL), 0.96 ($n = 209$) in the detection of medium paraprotein status (0.5–3.0 g/dL) and 1.00 ($n = 33$) in the detection of high paraprotein status (≥ 3.0 g/dL). In the external test set, the "Serum-Control" model achieved sensitivities of 0.90, 0.96, and 1.00 in samples with low, medium, and high quantified paraprotein status, respectively (Table 3).

Web service

We tested the web service by manually uploading 14 samples (2 with positive paraprotein status, 12 negative) from a randomly selected gel using HTTP POST request and retrieved the results. All 14 samples were scored the same by the web service as with the model as implemented in Python on a local machine.

Excel spreadsheet for automated paraprotein scoring

We tested the Excel spreadsheet by manually uploading 14 samples (2 with positive paraprotein status, 12 negative) from a randomly selected gel into the spreadsheet and calculating the results. All 14 samples were scored the same as with the model as implemented in Python on a local machine.

Discussion

Applicability to clinical pathology workflows

In our clinical laboratory, medical technologists first pre-interpret SPEP curves and subsequently send them to a pathologist for finalization. Our automated screening method tuned for high sensitivity aims to streamline this process. In this scenario, our novel method could apply a negative (no paraprotein) pre-interpretation to cases it determined did not have a paraprotein and send the case directly to a pathologist for finalization, allowing the medical technologist to focus their pre-interpretation efforts on the minority of cases that are likely to have a paraprotein (Table 4). Based on a 0.97 sensitivity in the external test set, this process would ensure identification of at least 97% of samples with a paraprotein, in order to facilitate appropriate downstream diagnosis and management.

In balance with the goal of catching positive cases with a screening method, is the goal of time savings associated with avoiding pre-interpretation for cases unlikely to be positive. Based on our external test set, which is sampled from a clinical laboratory at a major tertiary care center in which many samples are run for patients in post-treatment surveillance, about 65% of our SPEPs are negative, the model has a specificity of 0.29 and positive-predictive value of 0.42, representing a 20.0% time savings (Supplemental Data, Table 2). However, in a testing environment servicing the representative overall population, a majority of paraproteins identified would represent monoclonal gammopathy of unknown significance, which has a prevalence of 3.2%.²⁹ In a hypothetical scenario involving a paraprotein prevalence of 3.2%, our model would result in a 28.2% time savings (Supplemental Data, Table 3). Please note that the time saved depends on the prevalence of positive cases, and in the general community, there would be a fewer percentage of samples that have a paraprotein compared to our data. Furthermore, we defined time savings in regards to percentage of cases that could be screened out, and we do not quantify time savings in regards to real-world activities in the laboratory, which may include transmission of data from Sebia machinery to middleware or laboratory information systems. Furthermore, we do not consider time needed to build, maintain, or test algorithms. Additionally, we do not consider further classification of samples without a paraprotein in the model (e.g., the model may not detect a paraprotein in the sample, but the sample may have hypogammaglobulinemia and thus not be considered a normal sample); implementation and validation of such analysis remain as future work. Additionally, we believe that in clinical practice, leveraging our model in conjunction with human intervention would lead to the best outcomes in terms of identifying samples with paraproteins. For example, this may be necessary in cases where there is a paraprotein outside of the gamma region (e.g., beta region) because our method specifically only analyzes the gamma region for a paraprotein. In such cases, the model would be adopted into the pathologist's workflow and the model would likely identify the sample as negative for paraprotein. However, the sample would eventually be sent to a pathologist who would analyze the entire SPEP curve and can thus identify the paraprotein outside of the gamma region.

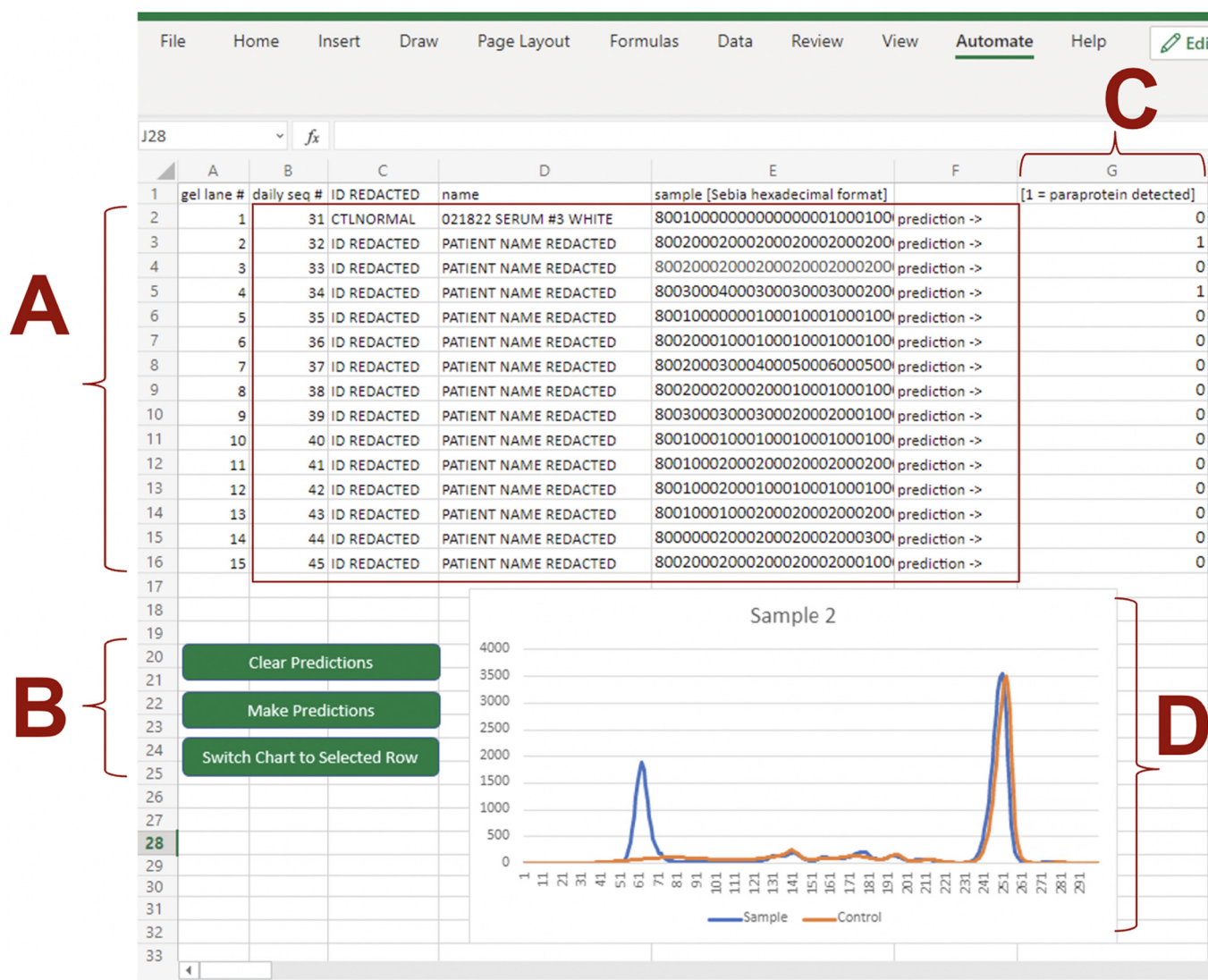


Fig. 6. Screenshot of the Excel spreadsheet macro. The user provides input in the form of tabular data (columns B–F in spreadsheet) copied from the database on the Sebia HYDRASYS system (A). The user then can call the macro functions by pressing the green buttons (B) for “Make Predictions” (sends GET request to web service for each row which returns a score (1 = paraprotein detected, 0 = paraprotein not detected), which is populated in column G (C), “Clear Predictions” (clears the cells for predictions in column G before the user replaces the copied data from Sebia HYDRASYS system), or “Switch Chart to Selected View” (changes the plot (D) to the sample and control curves associated with the row that the cursor is currently hovered over). The macro assumes that the normal serum control is in the first gel lane.

We reported strong performance of the model in both the representative test set as well as the external test set. The final “Serum–Control” model attained AUC scores of 0.90 in both the representative and external test sets, indicating that the model is robust enough to maintain the ability to separate positive and negative samples in a population of samples that differs from that on which the model was calibrated. Meanwhile the sensitivity of the model was higher in the external test set (0.97 vs 0.90 in the representative test set), while positive-predictive value was lower (0.42 vs 0.64 in the representative test set). Note that the representative test set and the external test set have different proportions of paraprotein-positive and paraprotein-negative samples. The representative test set consists of 50% positive samples (55/110) and the external test set consists of 35% positive samples (458/1321). We believe the higher sensitivity and lower specificity in the external test set are due to the operating point chosen to maximize likelihood ratio and can be changed in situations where a higher specificity may be desired. We believe the lower prevalence of positive samples in the external test set correlates with the lower positive-predictive value. We created such an external test set in order to rapidly test a large

number of samples en masse, as opposed to the representative test set whose purpose was to test the model on a set of samples that are representative of the samples that were used to calibrate the model.

Performance across different paraprotein levels is important to understand as laboratories may consider implementation of automated methods in practice in diagnostic workflows. When stratified across low, medium, or high paraprotein concentrations, strong performance persists across all strata when the model is scored on the external test set. The model calibrated to maximize likelihood ratio given a minimum sensitivity of 0.90, attained sensitivities of 0.90, 0.96, and 1.0 in samples with low, medium, and high quantified paraprotein concentrations, respectively when scored on the external test set. The relatively lower sensitivity in samples with low paraprotein levels is likely due to the fact that samples with low paraprotein levels often do not exhibit obvious spikes with high spatial frequency in the gamma region.

There are other previously proposed strategies for paraprotein detection in the literature, most of which are based on neural networks. We have included in the Supplemental Data a comparison of performance

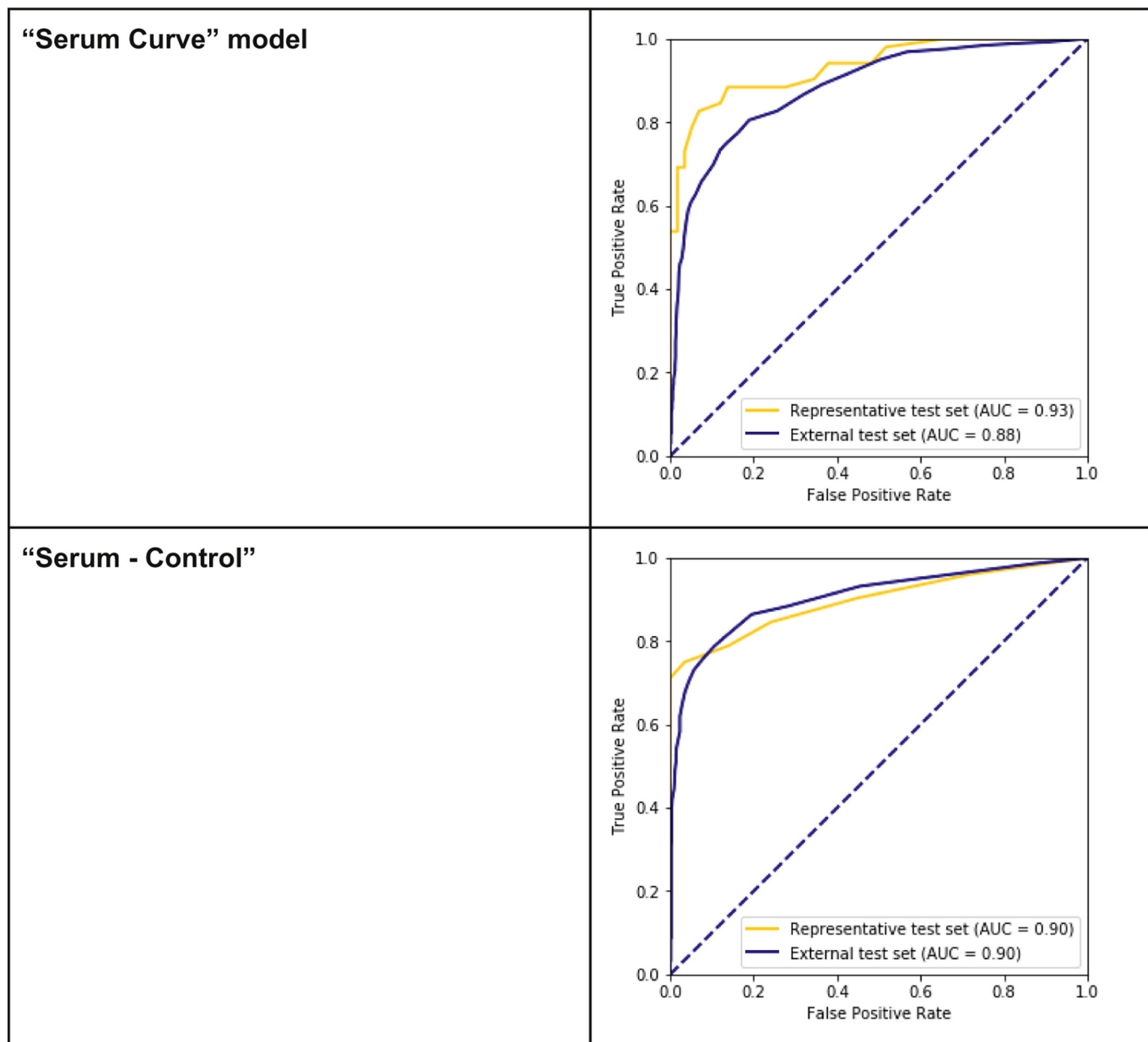


Fig. 7. Receiver operating characteristic curve for model scored on representative and external test sets.

metrics as reported by each of these methods (Supplemental Data, Table 4). Our method is the only one based on spatial frequency analysis, and the only one that has been implemented and is accessible entirely through open source tools. The only other published method with accessible code online⁷ requires data to be provided in the proprietary Microsoft Database format (.MDB), rather than in numerical array format (e.g., in the form of CSV files) in order to score samples. On the other hand, all other reported past works used approaches based on neural networks. Our method reported sensitivity of 0.97 in the external test set, which lies within 1% of other reported methods, although our method can be tuned to higher or lower sensitivity based on other desired goals such as specificity or time savings. Our method achieves AUC of 0.90 in the external test set; the only other methodology reporting AUC had an AUC of 0.99, although their methodology required training a large neural network on a large dataset of over 162 000 samples, requiring extensive computing resources (4 Tesla V100 GPUs).⁷ Other methods report specificity between 0.79 and 0.99, while our method achieves specificities of 0.55 in the representative test set and

0.29 in the external test set. Despite the trade-off in AUC or specificity, we maintain the stance that our method provides strong AUC performance while meeting sensitivity requirements of 0.90, without extensive computing resources for model training, and can be built and accessed publicly, completely with open source tools.

It is important to note that our method provides a lightweight implementation for paraprotein screening which can be utilized in resource-poor settings. Our implementation can be performed on a laptop computer and does not require expensive computing resources such as graphical processing units (we performed all of our model building and evaluation on a laptop computer with integrated graphics, and we did not require separate graphical processing units). Furthermore, our implementation is done with standard Python libraries and does not require extensive setup and configuration of deep learning libraries (e.g., TensorFlow, PyTorch), which can be considered a bottleneck to performing an automated screening process. Furthermore, our model is intended to be used for both calibration (learning the model parameters on the user's particular cohort, which varies between lab settings) as well as scoring (applying a trained model to one's own data). When using a

Table 2

Performance of model using the patient serum densitometry curve and the difference between the patient serum and control densitometry curves as input.

	Model with maximum likelihood ratio and minimum sensitivity 90%	
	Representative test set	External test set
<i>“Serum Curve” model</i>		
AUC	0.93	0.88
Sensitivity	0.90	0.92
Positive-predictive value	0.70	0.54
Specificity	0.66	0.58
F1 score	0.79	0.68
Accuracy	0.77	0.70
<i>“Serum–Control” model</i>		
AUC	0.90	0.90
Sensitivity	0.90	0.97
Positive-predictive value	0.64	0.42
Specificity	0.55	0.29
F1 score	0.75	0.59
Accuracy	0.72	0.52

Table 3

Sensitivity of the model calibrated to likelihood ratio (model with highest likelihood ratio while maintaining a minimum of 90% sensitivity) across quantified paraprotein status, evaluated in the external test set.

Quantified paraprotein status	n	Sensitivity (correct / n)	
		Baseline “serum curve” model	“Serum–Control” model
Low: 0.1–0.5 g/dL	216	0.86 (185/216)	0.90 (194/216)
Medium: 0.5–3.0 g/dL	209	0.96 (200/209)	0.96 (200/209)
High: ≥ 3.0 g/dL	33	1.00 (33/33)	1.00 (33/33)

computational method for paraprotein screening such as ours, we believe there should be a strong preference for using a model that was calibrated on data representative of samples that are to be scored. Thus, optimally a laboratory should be able to reliably calibrate and score models. However, it is also possible for a laboratory to use a model that is hosted remotely (and trained on data from another laboratory) to score their own samples. Nonetheless, in contrast with other methods requiring extensive resources for training, our model is lightweight enough to allow for both to be performed by a laboratory (rather than only using the model for scoring).

Taken together, our method can be particularly impactful in the initial screening process for paraproteins, which is often a tedious manual process. Speeding up the initial screening could in turn free up resources for more technically complex tasks in the clinical laboratory.

Limitations

Model calibration considerations of the current model result in performance trade-offs that should be considered individually within a laboratory’s testing context. Generally, as calibration is shifted towards a preference for higher sensitivity, a lower specificity results. The “Serum–Control” model, which was calibrated to maximize likelihood ratio while meeting a minimum sensitivity of 0.90, had an operating point with a

Table 4

A representation of potential time savings under various hypothetical scenarios in which the model could be used in clinical workflow, compared to a status quo without any computational model used. When using the model and when the model detects a sample is likely to be negative, the sample would be sent directly to the pathologist for interpretation, and pre-interpretation would not be completed by the technologist.

	Technologist pre-interpretation	Pathologist interpretation	Final result to be reported
Status quo (no computational model)	✓	✓	✓
Using our model: If sample is not likely negative	✓	✓	✓
Using our model: If sample likely negative	[skip - time savings]	✓	✓

sensitivity of 0.90 and specificity 0.55 in the representative test set. Had the model been calibrated to a minimum 0.95 sensitivity, the operating point would have had a sensitivity of 0.96 and specificity of 0.28 on the representative test set (Supplemental Data, “Calibration of model to sensitivity”). Outside of our baseline “Serum Curve” model, we do not believe that there is an appropriate baseline comparator for our model in terms of methodology or computing infrastructure because: (1) the goal of our study was to develop a signal processing based method that can be easily calibrated without extensive computational resources; (2) previous models proposed for paraprotein identification are complex models based on neural networks, which are a different domain of methodology and typically perform optimally when trained with large datasets, thus requiring more time or extensive computational resources such as graphical processing units.

Calibration may need to be done separately in situations with distinct testing populations, clinical settings, or data acquisition systems. For this analysis, all data sources acquired SPEP data using phoresis devices and software from Sebia. Further testing of the method using data sources from other acquisition devices (e.g., Helena Laboratories (Beaumont, TX, USA)) may aid efforts in generalizing the method to the broader range of clinical pathology and laboratory medicine environments.

Additionally, our model validation set included samples that were positive for paraproteins, but the paraproteins were located outside of the isolated gamma region. When removing such samples from the validation set, performance of the model exhibited an AUC improvement from 0.90 to 0.92 in the representative test set, 0.90 to 0.91 in the external test set (Supplemental Data, Table 1, Figs. 1, 2). Note that we were unable to programmatically isolate samples where paraproteins were strictly inside the gamma region, since this information is not readily available in structured data fields in the LIS database and thus not able to be queried. Thus, the entire validation and test sets included these samples. However, since the set of samples to be included in the study were chosen at random, they are representative of the overall population of samples in our LIS.

Our method is meant to be based on the electropherogram alone, and thus by design we did not include other features of the patient (e.g., demographics, other diagnoses) as input data to the method. Furthermore, samples may have abnormalities outside of paraprotein spikes, such as the presence of hypogammaglobulinemia. Some samples may actually have a monoclonal antibody, or a free light chain which would typically be detected with immunofixation. Such abnormalities are not meant to be specifically detected by our model.

Error analysis on samples in the representative and external test sets missed by the model indicate that samples with either dull paraprotein peaks or paraprotein peaks outside of the gamma region can be missed (Supplemental Data, Figs. 3–6). When the model was tuned to sensitivity, the “Serum–Control” model missed 1 positive paraprotein case out of 52 total positive cases in the representative test set, representing 1.9% of total positive cases missed (Supplemental Data, Table 3). In the external test set, the “Serum–Control” model missed 5 out of 458 positive cases, representing 1.1% of total positive cases missed (Supplemental Data, Fig. 4). When the model was tuned to likelihood ratio, the “Serum–Control” model missed 4 positive paraprotein cases out of 52 total positive cases in the representative test set, representing 7.7% of total positive cases missed (Supplemental Data, Fig. 5). In the external test set, the “Serum–Control” model missed 44 out of 458 positive cases, representing 9.6% of total positive cases missed (Supplemental Data, Fig. 6). Paraprotein peaks that are located

in the mid-gamma region or cathodal gamma regions are more likely to be picked up by the model, given that they exhibit enough spatial frequency. This performance is more due to the design of the model, which specifically isolates the gamma region before computing the spatial frequency spectrum. Also, the model was not intended to analyze the beta fraction. Any sample with a paraprotein in the beta fraction but not the gamma fraction, would have been predicted to be “negative” by the model due to the fact that the model only looks at the gamma fraction. Hypothetically, if these samples were to be scored by the model and the paraprotein was outside of the gamma region, the existence of a paraprotein would still need to be detected by the pathologist manually. In a real-life workflow, assuming the pathologist identifies paraproteins outside of the gamma region perfectly, these would eventually be interpreted as positive for a paraprotein using the combination of the computational model and manual curation. While we did not manually review samples for hypergammaglobulinemia, which typically exhibit dull peaks, our laboratory does use a hypergammaglobulinemia control in each gel we run. When testing the model with the web service on all samples on a randomly chosen gel, the model correctly classified the hypergammaglobulinemia control as negative for a paraprotein peak.

Future work

There are several opportunities for future work. First, the technique can be applied to immunofixation gels, where there are separate lanes for IgG, IgA, and IgM heavy chains as well as kappa and lambda light chains. Second, while spatial frequency analysis is an appropriate method for SPEP signals, additional signal processing techniques may be explored. For example, techniques such as wavelet analysis³⁰ and peak detection methods^{31,32} can potentially be leveraged. Furthermore, statistical characteristics of the signal, such as shape factor, kurtosis, and skewness, as well impulsive metrics such as peak value, impulse factor, and crest factor, can potentially yield insight in the context of screening for paraproteins.^{33,34} Third, although more cumbersome, methods that incorporate machine learning models including usage of other data domains like clinical lab values can be considered, with the caveat that feasibility depends on the laboratory having reliable access to these extra data domains. For example, this may involve a tiered method that first uses a spatial frequency-based screen for paraproteins, and then applies a machine learning model using clinical lab values to any sample that is predicted to be negative by the spatial frequency model. Such an approach may ultimately improve sensitivity. Additionally, natural language processing-based methods may be used to extract features from free-form text in clinician interpretation notes, which can potentially be useful for extraction of properties such as hypogammaglobulinemia. Future work may also consider extensions of the model to improve sensitivity in samples where the paraprotein is outside of the gamma region, or where the paraprotein is dull rather than sharp in appearance. Furthermore, extensions can be made to the model or the application programming interface (API) to warn users about potential variations of samples without paraproteins that may still warrant further evaluation (e.g., samples with hypogammaglobulinemia). Finally, additional avenues of future work may include the development of web applications or mobile applications to aid in dissemination and adoption of the method. For example, users may elect to clone the web service and host it on their own servers, or modify the application programming interface (API) for the web service to train their own model.

Conclusion

We have developed a paraprotein screening method using analysis of the spatial frequency domain of the serum protein electrophoresis curve, which achieved AUC of 0.90, positive-predictive value 0.64, specificity 0.55, F1-score 0.75, and accuracy of 0.72 while maintaining sensitivity of 0.90 (representative test set). Future work should focus on applying the technique to immunofixation gels, incorporating other data domains such as clinical laboratory values, and exploring lightweight machine learning-

based strategies that do not require extensive computational resources (e.g., GPUs) for model training (e.g., linear regression, logistic regression).

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.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jpi.2022.100128>.

References

- Rajkumar SV, Dimopoulos MA, Palumbo A, Blade J, Merlini G, Mateos M-V, et al. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *The Lancet Oncology* 2014;15(12):e538–e548. [https://doi.org/10.1016/S1470-2045\(14\)70442-5](https://doi.org/10.1016/S1470-2045(14)70442-5).
- O’Connell TX, Horita TJ, Kasravi B. Understanding and interpreting serum protein electrophoresis. *Am Fam Physician* 2005;71(1):105–112.
- Keren DF, Bocsi G, Billman BL, Etzell J, Faix JD, Kumar S, et al. Laboratory detection and initial diagnosis of monoclonal gammopathies: guideline from the College of American Pathologists in Collaboration with the American Association for Clinical Chemistry and the American Society for Clinical Pathology. *Archives of Pathology & Laboratory Medicine* 2021. <https://doi.org/10.5858/arpa.2020-0794-CP>.
- Kratzer MA, Ivandic B, Fateh-Moghadam A. Neuronal network analysis of serum electrophoresis. *Journal of Clinical Pathology* 1992;45(7):612–615. <https://doi.org/10.1136/jcp.45.7.612>.
- Männer GA, Schweiger CR, Söregi G, Pohl AL. Detection of monoclonal gammopathies in serum electrophoresis by neural networks. *Clinical Chemistry* 1993;39(9):1984–1985. <https://doi.org/10.1093/clinchem/39.9.1984>.
- Ognibene A, Motta R, Caldini A, Terreni A, Dea ED, Fabris M, et al. Artificial neural network-based algorithm for the evaluation of serum protein capillary electrophoresis. *Clinical Chemistry and Laboratory Medicine* 2004;42(12). <https://doi.org/10.1515/CCLM.2004.271>.
- Chabrun F, Dieu X, Ferre M, Gaillard O, Mery A, Chao de la Barca JM, et al. Achieving expert-level interpretation of serum protein electrophoresis through deep learning driven by human reasoning. *Clinical Chemistry* 2021;67(10):1406–1414. <https://doi.org/10.1093/clinchem/hvab133>.
- Borrillo F, Infusino I, Birindelli S, Panteghini M. Use of Neurosoft expert system improves turnaround time in a laboratory section specialized in protein diagnostics: a two-year experience. *Clinical Chemistry and Laboratory Medicine (CCLM)* 2021;59(9):e367–e369. <https://doi.org/10.1515/cclm-2021-0146>.
- Altinier S, Sarti L, Varagnolo M, Zaninotto M, Maggini M, Plebani M. An expert system for the classification of serum protein electrophoresis patterns. *Clinical Chemistry and Laboratory Medicine* 2008;46(10):1458–1463. <https://doi.org/10.1515/CCLM.2008.284>.
- Allen Broughton S, Bryan K. *Discrete Fourier Analysis and Wavelets: Applications to Signal and Image Processing*. John Wiley & Sons. 2018.
- Clifford GD, Azuaje F, McSharry P. *Advanced Methods and Tools for ECG Data Analysis*. Boston: Artech House Publishing. 2006.
- Sanei S, Chambers JA. *EEG Signal Processing*. John Wiley & Sons. 2013.
- Tereshchenko LG, Josephson ME. Frequency content and characteristics of ventricular conduction. *Journal of Electrocardiology* 2015;48(6):933–937. <https://doi.org/10.1016/j.jelectrocard.2015.08.034>.
- Gothwal H, Kedawat S, Kumar R. Cardiac arrhythmias detection in an ECG beat signal using fast fourier transform and artificial neural network. *Journal of Biomedical Science and Engineering* 2011;4(4):289. <https://doi.org/10.4236/jbise.2011.44039>.
- Polat K, Güneş S. Classification of epileptiform EEG using a hybrid system based on decision tree classifier and fast Fourier transform. *Applied Mathematics and Computation* 2007;187(2):1017–1026. <https://doi.org/10.1016/j.amc.2006.09.022>.
- PHORESIS Imaging System Instruction Manual. Sebia Electrophoresis. 2003:88–89.
- Cooley JW, Tukey JW. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation* 1965;19(90):297–301.
- Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical recipes: the art of scientific computing. *Analitica Chimica Acta* 1987;199:293–294. [https://doi.org/10.1016/S0003-2670\(00\)82860-3](https://doi.org/10.1016/S0003-2670(00)82860-3).
- Bressert E. *SciPy and NumPy: An Overview for Developers*. O’Reilly Media, Inc.. 2012
- McKinney W, Others. pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing* 14; 2011.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 2011;12(85):2825–2830.

22. Kyle RA, Therneau TM, Rajkumar SV, Larson DR, Plevak MF, Offord JR, et al. Prevalence of monoclonal gammopathy of undetermined significance. *The New England Journal of Medicine* 2006;354(13):1362–1369. <https://doi.org/10.1056/NEJMoa054494>.
23. Daubechies I. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory* 1990;36:961-1005.
24. Du P, Kibbe WA, Lin SM. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* 2006;22:2059–2065. <https://doi.org/10.1093/bioinformatics/btl355>.
25. Manikandan MS, Soman KP. A novel method for detecting R-peaks in electrocardiogram (ECG) signal. *Biomedical Signal Processing and Control* 2012;7(2):118–128. <https://doi.org/10.1016/j.bspc.2011.03.004>.
26. Smith SW. *The Scientist and Engineer's Guide to Digital Signal Processing*. California Technical Pub. 1997.
27. Smith S. *Digital Signal Processing: A Practical Guide for Engineers and Scientists*. Newnes. 2003.