EURASIAN PHYSICAL TECHNICAL JOURNAL

2026, Volume 23, No. 2 (56)

https://doi.org/10.31489/2026N2/156-164

Received: 27/01/2026 Revised: 07/06/2026 Accepted: 26/06/2026 Published online: 30/06/2026

Original Research Article

UDC: 524.3

VARIABLE STARS IN ARCHIVAL SCHMIDT CAMERA OBSERVATIONS: CROSS-MATCHING AND ML CLASSIFICATION

Izmailova I.*, Shomshekova S., Umirbayeva A., Aktay L.

Fesenkov Astrophysical Institute, Almaty, Kazakhstan
*Corresponding author: izmailova@fai.kz

Abstract. *Digitized archival astronomical observations provide a valuable basis for the study of variable stars and their long-term behavior. This work presents an analysis of variable stars based on archival photometric data obtained with a Schmidt camera at the Fesenkov Astrophysical Institute during the period from 1960 to 1989. The detected objects were cross-matched with existing catalogs of variable stars and modern astrometric and photometric databases, allowing the compilation of an extended set of stellar parameters that had not previously been combined within a single dataset. To complement incomplete catalog information, machine learning methods were applied to classify stellar variability types and spectral classes using available astrometric and photometric features. Several classification models were tested, and the most stable and accurate approach was selected for further analysis. The resulting catalog integrates archival observations, catalog cross-identification results, and predicted stellar characteristics, providing an extended resource for studies of stellar variability and long-term photometric evolution.*

Keywords: archival astronomical data, variable stars, catalog cross-matching, machine learning, photometric observations.

1. Introduction

Digitized archival astronomical data provide a unique opportunity to study long-term variability and evolution of celestial objects. While modern surveys such as Pan-STARRS [1], Gaia [2], and ZTF [3] deliver high-precision measurements over limited time spans, archival observations extend time series to several decades, which is crucial for investigating slowly evolving phenomena [4,5]. Combining archival and modern data enables the detection of long-term luminosity changes and trends that are not evident from contemporary surveys alone, particularly for variable stars. Recent advances in image processing and automation have significantly improved object identification, catalog expansion, and classification accuracy [6]. The development of virtual observatory infrastructures has facilitated access to digitized photographic plates, their cross-matching with modern catalogs, and large-scale automated analysis [7]. Previous studies have demonstrated that scanned plates allow reliable photometry [8] and that historical catalogs can be converted into machine-readable formats suitable for modern workflows [9].

In this work, we analyze archival Schmidt camera observations[1] obtained at the Fesenkov Astrophysical Institute between 1960 and 1989 and published through the Kazakhstani National Virtual Observatory. The data are cross-matched with the General Catalogue of Variable Stars [10], Gaia Data Release 3, and TESS

[1] https://vo.fai.kz/obs_data.php?path=/schmidt_telescope_lc/q/web/form

[11], enabling the refinement of stellar parameters such as galactic height, variability amplitude, and spatial distribution. To complement incomplete catalog information, machine learning methods are applied to predict spectral and variability types. The resulting catalog integrates archival photometry with astrometric, photometric, and physical parameters from modern surveys and includes information on the availability of archival observations. The catalog is published through the virtual observatory framework and is accessible via standard protocols, including the Table Access Protocol [12] and the Simple Cone Search [13].

2. Materials and Methods

This study includes several stages, from processing archival photometric data to the application of machine learning models for stellar classification. The overall workflow of the analysis is shown in Figure 1.

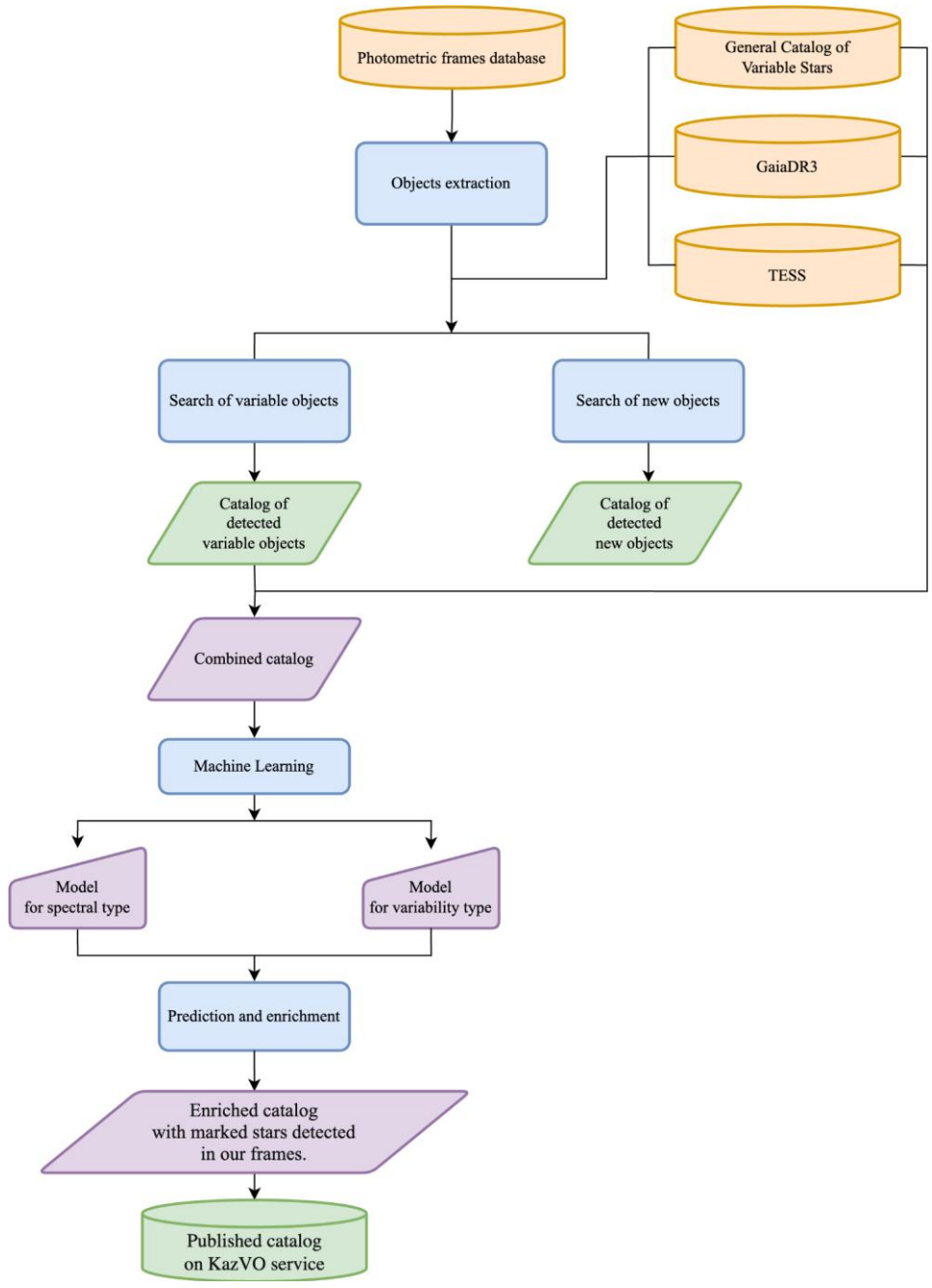

Fig.1. Workflow of the study from archival data digitization to machine learning model training.

2.1. Photometric Data and Coverage

The analysis is based on archival photometric data obtained with the Schmidt camera and published through the Kazakhstani National Virtual Observatory. The digitization and publication procedures have been described in detail in previous studies [14,15]. The spatial coverage of the archival observations is illustrated in Figure 2. The archive contains 1270 Schmidt camera frames covering approximately 74.6% of the sky. The digitized frames represent raw optical density scans and are not photometrically calibrated. Although a limited number of calibration plates are available, their coverage is insufficient to ensure uniform calibration across the dataset. Therefore, photometric measurements derived directly from the plates were not used in this study.

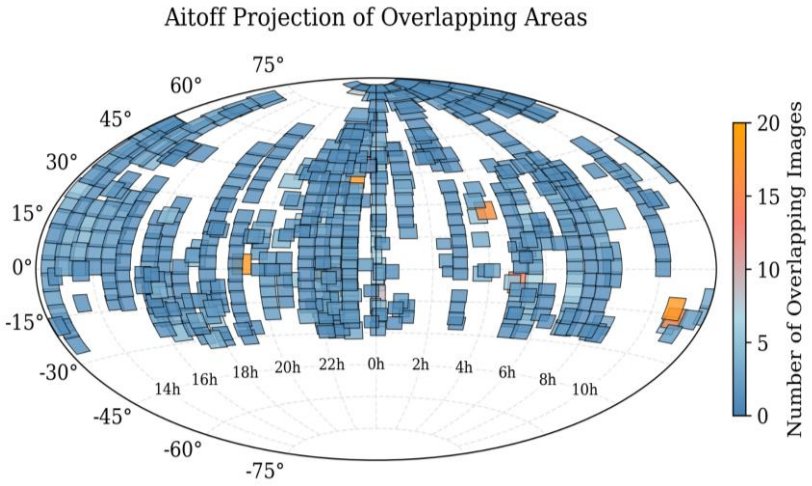

Fig.2. Coverage of photometric observations from the Schmidt camera. Rectangles show observed sky regions; color indicates the number of frames.

Despite this limitation, the extensive sky coverage of the archival material allows reliable coordinate-based cross-identification with modern catalogs and supports the integration of historical observations into contemporary datasets.

2.2. Source Extraction

Source extraction was performed automatically using a custom Python-based pipeline[2] built on the SEP library [16], which implements the Source Extractor algorithm [17]. Background subtraction parameters were optimized by minimizing residual noise and brightness variations, after which background-corrected images were saved in FITS format. Object detection was carried out in image segments to reduce memory usage. Detection thresholds were defined relative to the background, with a minimum object area of nine pixels. Overlapping sources were separated using multi-level segmentation.

For each detected object, coordinates, fluxes, magnitudes, and associated uncertainties were extracted and saved in tabular format. The extraction results were visually validated by overlaying detected sources on the original images. The pipeline supports batch processing, multithreading, and command-line execution, enabling fully automated source extraction for large archival datasets.

2.3. Variable Stars Search

After source extraction and coordinate determination, the detected objects were cross-matched with catalogs of variable stars. The initial step involved transforming the extracted coordinates into a unified reference system compatible with the General Catalogue of Variable Stars. To determine an appropriate matching radius, a test subset comprising the brightest detected objects was cross-matched with the Gaia catalog using a nearest-neighbor approach. The search radius was gradually increased until stable matches were obtained for all test objects, and the resulting coordinate discrepancies were analyzed. Outliers associated with optical distortions were excluded, and the final matching radius was set to 21.75 arcseconds,

[2] https://github.com/ill-i/variable_search (accessed 27 January 2026)

corresponding to a 5-sigma deviation from the mean positional offset distribution as illustrated in the left panel of Figure 3. This value provided a balance between positional tolerance and matching reliability.

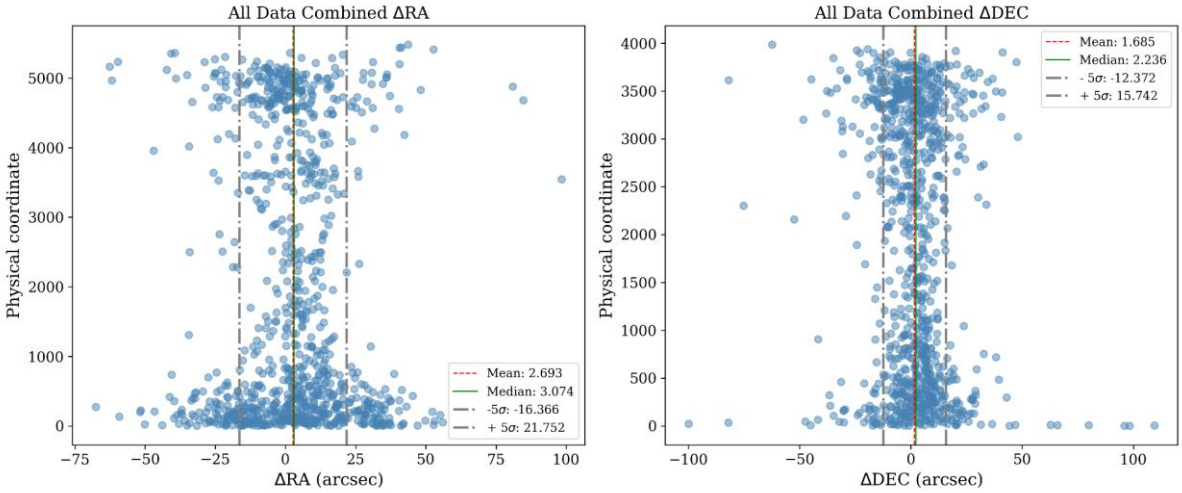

Fig.3. Distribution of coordinate discrepancies in RA (left) and Dec (right) between detected objects and Gaia data. Mean and median values, as well as 5-sigma boundaries, are marked with colored lines.

Using the derived search radius, the full dataset was cross-matched with the General Catalogue of Variable Stars. To improve computational efficiency, the dataset was processed in subsets. The resulting matches were compiled into a unified table containing positional and classification information from both archival detections and reference catalogs.

2.4. Search for New Objects

In addition to identifying known variable stars, a search for objects absent from the Gaia Data Release 3 catalog was performed to assess the consistency of the archival dataset and to identify potential new sources. For this purpose, archival object coordinates were transformed into Cartesian space and indexed using a nearest-neighbor search algorithm. An object was classified as new if no Gaia counterpart was found within the matching radius determined in the previous subsection. To reduce computational complexity, Gaia sources were pre-filtered using the spatial boundaries of individual archival frames. Candidate objects were recorded together with their positional information and frame identifiers. This approach does not account for brightness variability and is therefore not suitable for detecting short-term transient phenomena.

2.5. Machine Learning

To extend the characterization of the identified variable stars, supervised machine learning techniques were applied to classify stellar variability types and spectral classes using parameters derived from modern astronomical catalogs. Since variability and spectral labels are available for different subsets of objects, two independent classification tasks were defined: variability type prediction and spectral classification. This approach allowed the use of the largest possible labeled datasets for each task while preserving classification reliability.

2.5.1 Data Preparation

The input features were compiled from three primary sources: the General Catalogue of Variable Stars, Gaia Data Release 3, and TESS. These catalogs provide variability and spectral labels, astrometric parameters, photometric measurements, color indices, and selected astrophysical and physical stellar properties. Galactic coordinates and galactocentric distances were computed from sky positions and distance estimates. The combined dataset was cleaned by removing invalid entries and converting numerical parameters into a unified format. To reduce scale variance, the variability period was log-transformed. Parameters derived directly from digitized photographic plates were not used for classification due to the absence of reliable photometric calibration; however, the presence of archival observations was retained as metadata.Variability types were grouped into six major classes (eruptive, pulsating, eclipsing, cataclysmic, rotating, and X-ray). Objects with multiple variability types were simplified to a single-label multiclass formulation by retaining only the primary variability type, while ambiguous entries were excluded. Evaluation metrics, including accuracy, were

computed using exact-match criteria for the dominant class. Spectral types from different catalogs were consolidated into broad classes, and uncertain classifications were omitted from model training.

2.5.2 Machine Learning Models

Machine learning models were trained separately for variability type classification and spectral classification using astrometric, photometric, and astrophysical features. Several algorithms were evaluated, including gradient boosting methods, ensemble tree models, and neural networks. The dataset was divided into training and test subsets using an 80/20 split. Missing values were not imputed, as they represent an intrinsic characteristic of observational astronomical data and may introduce systematic biases if interpolated. Hyperparameter optimization was performed using Bayesian optimization minimizing the logarithmic loss function. Hyperparameter tuning and feature selection were conducted strictly on the training subset using k-fold cross-validation. The 20% test subset was isolated and used exclusively for the final metric evaluation presented in Figures 5–6. Feature importance analysis was applied to reduce model complexity while preserving physically meaningful input parameters. The final set of features used in both classification tasks is summarized in Table 1.

Table 1. Input features used for machine learning classification.

Feature group	Parameters	Source
Astrometric parameters	Galactic longitude (LII), Galactic latitude (BII), Parallax (log), Total proper motion, Galactocentric distance (log)	GCVS, Gaia DR3
Photometric and color indices	G-band magnitude, BP–RP, BP–G, G–RP, TESS-band magnitude, Extinction (log), Color excess E(BP–RP)	Gaia DR3, TESS
Variability parameters	Period (log), Cycle period flag	GCVS
Stellar physical properties	Effective temperature (log), Surface gravity (log g), Metallicity [Fe/H], Radius (log), Mass, Stellar density (log), Luminosity (log), Distance (log)	Gaia DR3, TESS

2.6. Predictions and Catalog Enrichment

The optimized models were applied to predict variability types and spectral classes for objects lacking complete classifications in the combined catalog. The trained models were saved for reuse and applied without retraining to ensure reproducibility. As a result, the catalog was enriched with predicted classifications and augmented with metadata indicating the availability of archival observations for each object. The final dataset integrates information from archival observations, modern catalogs, and machine learning predictions. The enriched catalog was published through the Kazakhstani National Virtual Observatory[3] using the DaCHS tool [18] and the TAP service, enabling further analysis and cross-identification.

3. Results and Discussion

This section presents the main results of the archival data analysis and discusses their implications for stellar classification and catalog enrichment.

3.1. Identification of Variable Stars and Archival Coverage

Analysis of the archival Schmidt camera observations resulted in the identification of 20,961 variable objects. Figure 4 illustrates the spatial overlap between the observational frames and variable stars listed in the General Catalogue of Variable Stars. Approximately 36.1% of catalogued variable stars fall within the coverage of the archival observations, demonstrating the significant potential of the digitized archive for extending temporal baselines and refining variability studies.

A search for objects without counterparts in the Gaia Data Release 3 catalog did not reveal reliable transient candidates. This result is primarily attributed to the limited astrometric accuracy of the archival frames, with a positional uncertainty of approximately 21.75 arcseconds, which is insufficient for confident

[3] https://vo.fai.kz/

detection of unmatched sources given the sub-arcsecond precision of Gaia. Consequently, coordinate-based cross-identification proved suitable for catalog consistency checks but ineffective for transient detection.

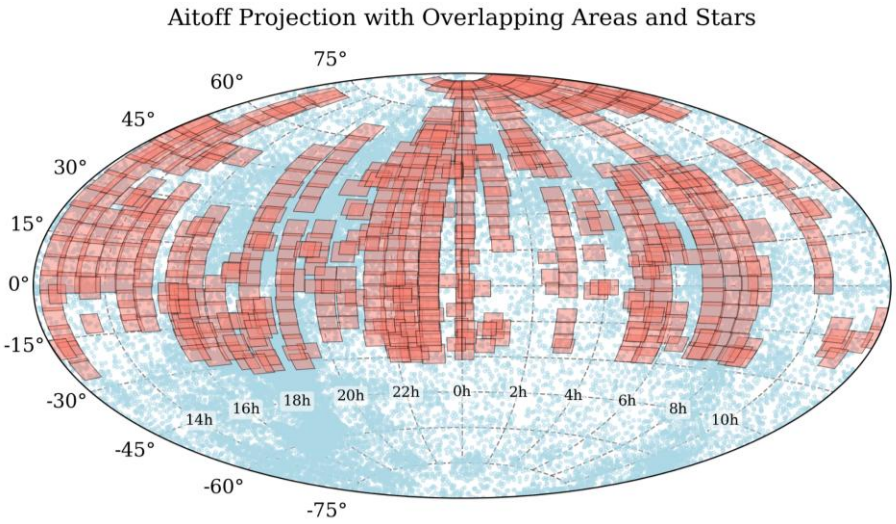

Fig.4. Distribution of variable stars from the GCVS catalog (blue dots) and observational frames obtained with the Schmidt camera (red rectangles).

3.2. Performance of Machine Learning Models

Machine learning methods were applied to classify stellar variability types and spectral classes using catalog-derived features. Among the evaluated machine learning models, XGBoost demonstrated the highest overall performance for both variability and spectral type classification tasks, particularly for rare classes. LightGBM showed comparable results for spectral classification, while Random Forest and neural network models exhibited reduced robustness and generalization capability. The comparative performance of the models for variability type classification is shown in Figure 5.

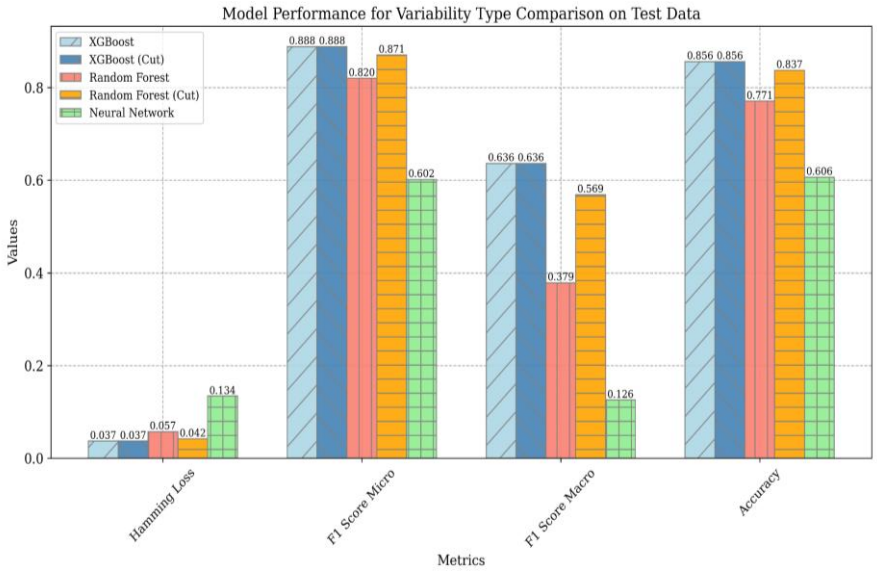

Fig.5. Comparison of machine learning model performance for variability type classification on test data.

As illustrated in Figure 5, feature selection based on individual importance scores led to a systematic decrease in classification performance, indicating the contribution of low-importance features through non-linear interactions. The results of spectral type classification are presented in Figure 6.

Experiments with feature selection indicated that removing parameters with low individual importance consistently degraded classification performance. This suggests that such features contribute through non-linear interactions, and therefore the complete feature set was retained in the final model configuration.

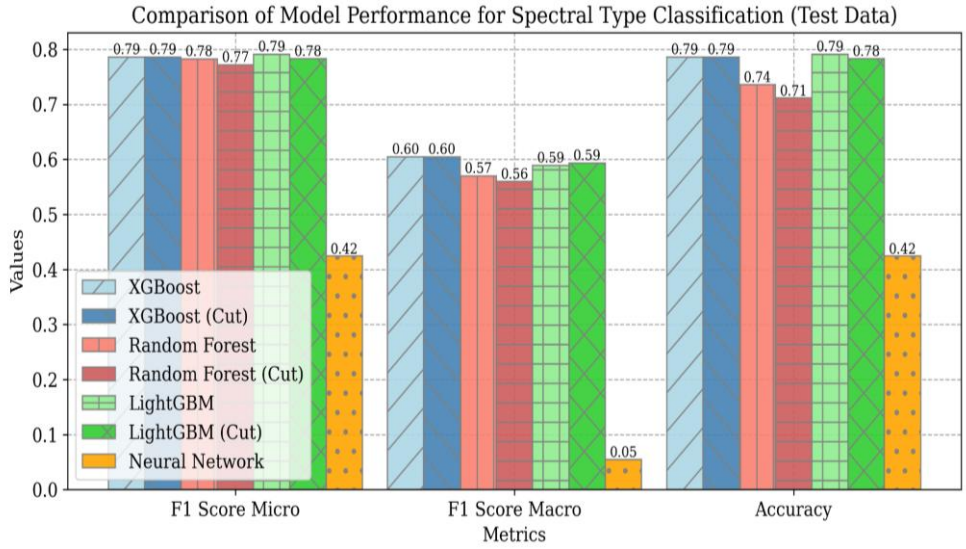

Fig.6. Comparison of machine learning model performance for spectral type classification on test data.

The main instances of model misclassification occur within regions of high interstellar extinction (the Galactic plane), which distorts the color indices, and among variable types with overlapping parameter profiles.

3.3. Catalog Enrichment and Validation

The optimized models were used to predict missing variability and spectral classifications in the combined catalog. As a result, spectral types were assigned to 40,980 stars with previously unknown classifications (99.95%), while variability types were predicted for 477 out of 538 unlabeled cases (88.66%).

Model reliability was assessed using standard performance metrics on held-out test data. The average accuracy of spectral type classification reached 78.9%, with a macro-averaged F1-score of 60%, indicating balanced performance across both common and rare classes. Variability type classification achieved an average accuracy of 86.5%, with micro- and macro-averaged F1-scores of 89.6% and 63.8%, respectively. Lower recall values for underrepresented classes, such as X-ray and rotating variables, reflect their intrinsic rarity in the training data. The spatial distribution of predicted classes follows expected galactic patterns, with late-type stars tracing the Galactic disk and rarer objects exhibiting broader distributions. This consistency supports the physical plausibility of the predictions and indicates the absence of systematic artifacts introduced by the models.

3.4. Publication and Applicability of the Catalog

The enriched catalog[4], integrating archival observations, modern survey data, and machine learning predictions, has been published through the Kazakhstani National Virtual Observatory in compliance with IVOA standards. It is accessible via standard data access protocols, enabling automated queries and integration into further studies of stellar variability and long-term photometric behavior. Future work may focus on incorporating brightness-based variability detection, multi-epoch observations, and probabilistic classification approaches to improve transient sensitivity and uncertainty estimation.

4. Conclusions

This work presents an integrated workflow for processing digitized archival photometric observations and combining them with modern astronomical catalogs, supplemented by machine learning–based

[4] https://vo.fai.kz/obs_data.php?path=/var_star_cat/q/scs/form

classification. Using archival Schmidt camera data obtained between 1960 and 1989, 20,961 variable stars were identified within the observational coverage, which spans approximately 75% of the sky. Cross-identification with contemporary catalogs enabled the extension of their astrometric, photometric, and astrophysical characterization beyond what is available from archival material alone.

The application of supervised machine learning methods allowed the enrichment of the catalog by predicting missing stellar parameters. Spectral types were assigned to 40,980 stars and variability types to 477 objects lacking prior classifications. The achieved classification performance is consistent with results reported in recent catalog-based studies, with average accuracies of 78.9% for spectral types and 86.5% for variability types. The spatial distribution of predicted rare classes, such as white dwarfs and chemically peculiar stars, follows expected Galactic trends, indicating that the models do not introduce artificial clustering or systematic biases. The analysis also demonstrates that coordinate-based cross-matching of archival data is effective for catalog consistency checks but insufficient for reliable transient detection, due to the limited astrometric accuracy and temporal sampling of photographic plate material. This finding is consistent with earlier studies and underscores the need for time-domain photometric information in transient searches.

The resulting catalog, integrating archival observations, modern survey data, and inferred classifications, has been published through the Kazakhstani National Virtual Observatory and made accessible via standard IVOA protocols. Its scientific novelty lies in the systematic integration of historical observations with contemporary catalogs and automated classification techniques. The catalog provides a practical resource for studies of long-term stellar variability, the statistical properties of rare stellar populations, and the planning of targeted follow-up observations. Future work may extend this approach by incorporating calibrated photometry, multi-epoch data, and probabilistic classification methods to further improve completeness and reliability.

Conflict of interest statement

The authors declare that they have no conflict of interest in relation to this research, whether financial, personal, authorship or otherwise, that could affect the research and its results presented in this paper.

CRediT author statement

Izmailova I.M.: Conceptualization, Methodology, Software, Visualization, Writing – Original Draft; **Umirbayeva A.Zh.:** Formal analysis, Validation, Writing – Review & Editing; **Shomshekova S.A.:** Supervision, Project administration, Funding acquisition**; Aktay L.:** Data curation, Investigation. The final manuscript was read and approved by all authors.

Statement on the use of Artificial Intelligence.

During the preparation of this manuscript, artificial intelligence tools were used solely for language editing and grammatical improvement. No AI tools were used to generate scientific content, analysis, results, or conclusions.

Data Availability Statement

The data that support the findings of this article are openly available via the Kazakhstan National Virtual Observatory (KazVO) platform and can be accessed using standard Table Access Protocol (TAP) services.

Funding

This work was supported by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan under Grant No. AP22784884.

Acknowledgements

The authors express their sincere gratitude to Dr. A. Serebryanskiy and Dr. V. Kim (Fesenkov Astrophysical Institute) for their valuable discussions and insightful comments, which contributed to the development of this work.

The authors also acknowledge the Vatican Observatory Summer School (VOSS 2023), where one of the authors gained essential knowledge and practical skills in machine learning that were instrumental for this study.

References

1 Chambers, K.C., Magnier, E.A., Metcalfe, N., Flewelling, H.A., Huber, M.E., Waters, C.Z., Denneau, L., Draper, P.W., Farrow, D., Finkbeiner, D.P., et al. (2016). *The Pan-STARRS1 surveys*. arXiv:1612.05560. https://doi.org/10.48550/arXiv.1612.05560

2 Gaia Collaboration, Prusti T., de Bruijne, J.H.J., Brown, A.G.A., Vallenari, A., Babusiaux, C., Bailer-Jones, C.A.L., Bastian, U., Biermann, M., Evans, D.W., et al. (2016). The Gaia mission. *Astronomy and Astrophysics*, 595, A1. https://doi.org/10.1051/0004-6361/201629272

3 Masci, F.J., Laher, R.R., Rusholme, B., Shupe, D.L., Groom, S., Surace, J., Jackson, E., Monkewitz, S., Beck, R., Flynn, D., et al. (2019). The Zwicky Transient Facility: Data processing, products, and archive. *Publications of the Astronomical Society of the Pacific*, 131(995), 018003. https://doi.org/10.1088/1538-3873/aae8ac

4 Jia, P., Yang, Z., Shang, Z., Yu, Y., & Zhao, J. (2023). Data processing pipeline for multiple-exposure photo-plate digital archives. *Publications of the Astronomical Society of Japan*, 75, 811–824. https://doi.org/10.1093/pasj/psad038

5 Grindlay, J., Tang, S., Los, E., Servillat, M. (2012). Opening the 100-year window for time-domain astronomy. *Proceedings of the IAU Symposium 285: New Horizons in Time-Domain Astronomy*. Cambridge University Press, 29–34. https://doi.org/10.1017/S1743921312000166

6 Kolesnikova, D.M., Sat, L.A., Sokolovsky, K.V., Antipin, S.V., Belinskii, A.A., Samus', N.N. (2010). New variable stars on digitized Moscow collection plates: The field of 66 Ophiuchi. *Astronomy Reports*, 54, 1000–1018. https://doi.org/10.1134/S1063772910110065

7 Shlyapnikov, A.A., Gorbunov, M.A., Gorbachev, M.A. (2020). Archives of CrAO spectral observations. Catalogues of objects and images. *Astronomical Archives Transactions*, 1, 23.

8 Sokolovsky, K.V., Zubareva, A.M., Kolesnikova, D.M., Samus, N.N., Antipin, S.V., Belinski, A.A. (2017). Accurate photometry with digitized photographic plates of the Moscow collection. *Proceedings of the IAU Symposium 339: Southern Horizons in Time-Domain Astronomy*. https://doi.org/10.48550/arXiv.1712.04672

9 Gorbunov, M.A., Shlyapnikov, A.A. (2017). Identification of stars and digital version of E.S. Brodskaya and V.F. *Shajn catalogue of 1958. Astrophysics – Instrumentation and Methods for Astrophysics*, 7 [in Russian] https://doi.org/10.48550/arXiv.1709.08113

10 Samus', N.N., Kazarovets, E.V., Durlevich, O.V., Kireeva, N.N., Pastukhova, E.N. (2017). General catalogue of variable stars: Version GCVS 5.1. *Astronomy Reports*, 61, 80–88. https://doi.org/10.1134/S1063772917010085

11 Ricker, G.R., Winn, J.N., Vanderspek, R., Latham, D.W., Bakos, G.A., Bean, J.L., Berta-Thompson, Z.K., Brown, T.M., Buchhave, L., Butler, N.R., et al. (2015). Transiting Exoplanet Survey Satellite (TESS). *Journal of Astronomical Telescopes, Instruments, and Systems*, 1, 014003. https://doi.org/10.1117/1.JATIS.1.1.014003

12 Dowler, P., Rixon, G., Tody, D., Demleitner, M. (2019). Table Access Protocol, Version 1.1. *IVOA Recommendation, Data Access Layer Working Group*. http://www.ivoa.net/documents/TAP/20190927

13 Plante, R., Williams, R., Hanisch, R., Szalay, A. (2008). Simple Cone Search, Version 1.03. *IVOA Recommendation, Data Access Layer Working Group*. https://doi.org/10.5479/ADS/bib/2008ivoa.specQ0222P

14 Shomshekova, S.A., Izmailova, I.M., Moshkina, S.G., Umirbayeva, A.Zh. (2022). Digitization of cometary photometric astroplates of the Fesenkov Astrophysical Institute. *Proceedings of the National Academy of Sciences of the Republic of Kazakhstan*, 1(341), 137–143. https://doi.org/10.32014/2022.2518-1483.143 [in Russian]

15 Shomshekova, S., Izmailova, I., Umirbayeva, A., Omarov, C. (2022). A method for digitization of archival astroplates of the Fesenkov Astrophysical Institute. *New Astronomy*, 97, 101881. https://doi.org/10.1016/j.newast.2022.101881

16 Barbary, K. (2016). *SEP: Source Extractor as a library. Journal of Open-Source Software*, 1, 58. https://doi.org/10.21105/joss.00058

17 Bertin, E., Arnouts, S. (1996). SExtractor: Software for source extraction. *Astronomy and Astrophysics Supplement Series*, 117, 393–404. https://doi.org/10.1051/aas:1996164

18 Demleitner, M., Neves, M.C., Rothmaier, F., Wambsganss, J. (2014). Virtual observatory publishing with DaCHS. *Astronomy and Computing*, 7, 27–36. https://doi.org/10.1016/j.ascom.2014.08.003

AUTHORS' INFORMATION

Izmailova, Ildana - M.Sc., Junior Researcher, Fesenkov Astrophysical Institute, Almaty, Kazakhstan; ScopusID: 57776936800; ORCID iD: 0000-0001-9878-0989; izmailova@fai.kz

Umirbayeva, Adel - M.Sc., Junior Researcher, Fesenkov Astrophysical Institute, Almaty, Kazakhstan; ScopusID: 57776936900 ; ORCID iD: 0000-0001-9339-4990; umirbayeva@fai.kz

Shomshekova, Saule - Ph.D., Leading Researcher, Fesenkov Astrophysical Institute, Almaty, Kazakhstan; ScopusID: 57208861899; ORCID iD: 0000-0002-9841-453X; shomshekova@fai.kz

Aktay, Laura – Barchelor, Engineer, Fesenkov Astrophysical Institute, Almaty, Kazakhstan; ORCID iD: 0009-0005-5862-4777; aktay@fai.kz