УДК 633.1:633

B.B. Mekecha, A.V. Gorbatov

Crop yield prediction in Ethiopia using gradient boosting regression

Nowadays, machine learning algorithms and methods are used in multiple areas of studies to achieve practical and productive solutions. Agriculture is one of the industries where the impact is significant, especially in the area of crop yield prediction and crop selection which is crucial for ensuring food security and improving agricultural practices. In a country like Ethiopia, where the economy is highly dependent on agriculture, and farming in particular, leveraging the powers of AI and machine learning is crucial. However, the use of these technologies in Ethiopian agriculture remains limited, mainly due to the lack of well-organized and digital datasets and lack of technological advancements.

The aim of this study is to increase the accuracy of crop yield prediction in Ethiopia and provide information that can help farmers and policymakers improve crop productivity. In this study, a crop yield prediction model was developed based on historical data that includes factors such as crop type, rainfall, temperature, Area cultivated, production, and pesticides.

Among the algorithms considered in this study, GradientBoostingRegressor achieved the highest value of the R-square – 90% compared to others which indicates its best predictive ability. However, the study also acknowledges the contextual advantages of other algorithms, highlighting the importance of selecting models that are appropriate for specific data sets and purposes. The accuracy and efficiency of agricultural planning and resource allocation in Ethiopia can be greatly improved by using machine learning techniques for crop production prediction.

Keywords: Crop yield, machine learning algorithms, food security, Ethiopia.

DOI: 10.21293/1818-0442-2024-27-3-125-129

Ethiopia's agricultural sector plays a vital role in its economy, accounting for 40% of the country’s gross domestic product (GDP), 80% of exports, and an estimated 75% of the country's workforce (employment) [1]. «Ethiopia's crop agriculture, mainly driven by small farms that cultivate cereals such as teff, wheat, maize, sorghum, and barley, and oats has significantly contributed to the total value added. Smallholders, who make up 96% of the total farmed land, primarily produce cereals for both consumption and sales. Due to the limited amount of land available for cultivation, especially in the highlands, future growth in cereal production will largely depend on yield improvements» [2].

Studies also showed that there is a significant annual fluctuation in cereal crop yield from 1994 to 2021, which underlines the significance of understanding these trends for sustainable agricultural planning and food security [3]. crop yield prediction using machine learning techniques is essential for effective agricultural planning and resource allocation in Ethiopia [4].

Traditional methods often lack accuracy and timeliness, prompting the exploration of advanced techniques such as machine learning (ML) [5]. Robots, sensors, drones, and algorithms can execute traditional agricultural tasks more quickly, such as weeding, pesticide application, irrigation, fertilizer recommendation, and soil nutrition and moisture status monitoring [6–8].

Our research builds on using machine learning algorithms and domain-specific data. By integrating support vector models (SVM), long short-term memory (LSTM), and recurrent neural networks (RNN), previous studies have shown promising results in predicting crop yield based on various factors such as water availability and fertilizer use [9]. In addition, the integration of geospatial technologies and the Internet of Things with machine learning algorithms has opened up new opportunities for real-time monitoring and decision making in agriculture [10]. Furthermore, recent studies have shown the potential of new crop yield forecasting methods specific to Ethiopian agriculture. An analogue approach to crop yield forecasting in the Upper Blue Nile Basin of Ethiopia, for example, uses historical soil moisture and crop yield data to achieve high forecast accuracy, addressing the limitations of traditional forecasting methods and providing valuable information for real-time seasonal forecasts [11].

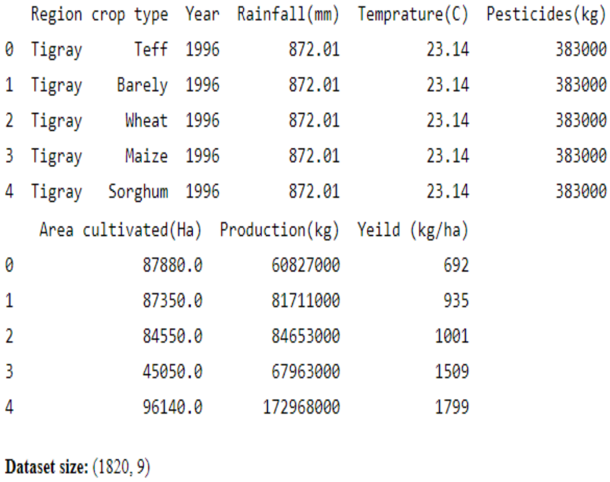
Moreover, advances in the use of remote sensing data and machine learning algorithms have enabled accurate prediction of agricultural losses due to drought in Ethiopia, offering important information for early intervention planning and improvement of existing early warning systems [12]. These developments highlight the importance of innovative approaches to crop yield forecasting and their potential to revolutionize agricultural management practices in Ethiopia and similar regions. This paper presents a customized approach to yield prediction focusing on seven major crops namely teff, barely, wheat, maize, sorghum, millet and oats in the national regional states of Ethiopia.

Material and methods

Historical crop yield data for the period 1996 to 2022 for nine regions and one federal level city administration (Dire Dawa) is collected from the Central Statistics Agency of Ethiopia (CSA) [13]. Climate data and pesticide information are from FAO and World Bank, respectively [14, 15]. The dataset is composed of 1,820 samples each with nine unique attributes. Using Jupyter Notebook as the platform. Figure 1 below shows an example of the dataset before preprocessing.

The initial data undergoes preprocessing to handle Null values, removing outliers, applying OneHotEncoder to categorical features, and normalizing features. After preprocessing, our dataset is reduced to 1,132 samples. To make model training and evaluation easier, the dataset is divided into training and testing subsets. The dataset is then subjected to a variety of Regression models. Gradient Boosting Regression, Random Forest Regression, Decision Tree Regression, Gaussian Process Regressor, Kneighbors Regressor and Linear Regression models are trained on a pre-processed dataset, taking into account factors specific to the Ethiopian agricultural context.

For effective training and testing of our model, we used an 80/20 ratio. In particular, 905 samples, or 80% of the dataset, were allocated for training the model, and 227 samples, or 20% of the dataset, were randomly selected for testing its performance. Each region crop combination was considered as a separate time series to ensure model training and testing reflected real-world conditions. Finally, the most accurate model is chosen by carefully assessing its ability to predict crop yields using the testing dataset. Figure 2 below clearly shows how the system components interact with each other, starting with preprocessing the data and ending with analyzing the results and choosing the best model. The proposed system is capable of determining crop yields.



```
  Region crop type  Year  Rainfall(mm)  Temprature(C)  Pesticides(kg)
0  Tigray      Teff  1996        872.01          23.14          383000
1  Tigray    Barely  1996        872.01          23.14          383000
2  Tigray     Wheat  1996        872.01          23.14          383000
3  Tigray     Maize  1996        872.01          23.14          383000
4  Tigray   Sorghum  1996        872.01          23.14          383000
   Area cultivated(Ha)  Production(kg)  Yeild (kg/ha)
0              87880.0        60827000            692
1              87350.0        81711000            935
2              84550.0        84653000           1001
3              45050.0        67963000           1509
4              96140.0       172968000           1799

Dataset size: (1820, 9)
```

Fig. 1. Sample Dataset

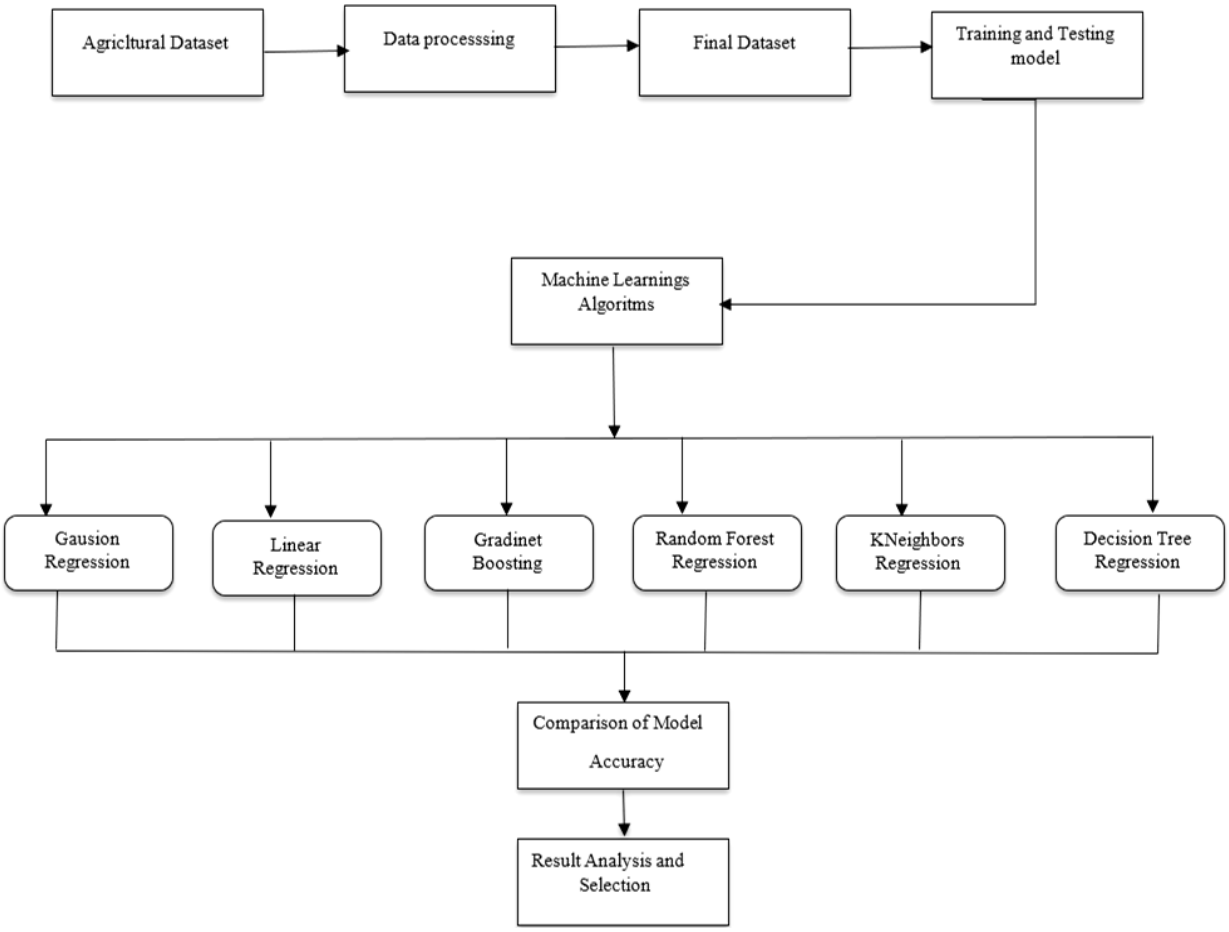


Fig. 2. Block Diagram of Proposed system

Results and Discussion

The effectiveness of each model is evaluated based on Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and the coefficient of determination (R-squared index) Values. As described in Table 1 among the evaluated models, GradientBoostingRegressor and DecisionTreeregressor demonstrate relatively higher R2 values compared to others, indicating higher prediction accuracy. In addition, both Gradient Boosting Regression and Decision Tree regression exhibit lower RMSE and MAE values, which on average indicate lower prediction errors.

While GradientBoostingRegressor achieves the highest R-squared score of 90%, indicating a best fit to the data compared to other models. DecisionTreeregressor follows with an R-squared score of 85%, which makes it the second best model. Linear Regression indicates a relatively high R-squared scores of 75%, outperforming the RandomForestRegressor with 68%. Meanwhile, the KNeighborsRegressor and GaussianProcessRegressor achieve R-squared scores of 66% and 51%, respectively, showing that they also perform reasonably well, though less effectively than the other models.

Table 1

Model Performance Comparison

Models	RMSE	MAE	(R^2) score(%)
GradientBoostingRegressor	0.04	0.03	0.90
RandomForest Regressor	0.08	0.06	0.68
DecisionTree regressor	0.05	0.03	0.85
GaussianProcess Regressor	0.09	0.06	0.51
KNeighborsRegressor	0.08	0.05	0.66
Linear Regression	0.07	0.05	0.75

These results provide valuable information on the comparative effectiveness of various regression models, allowing stakeholders to select the most appropriate model for predicting crop yields based on specific requirements and preferences.

Table 2 summarizes the hyperparameters chosen for the GradientboostingRegressor model. The chosen configurations, such as 0.1 learnig_rate and 200 n_estimators, were carefully chosen to strike a balance between model complexity and performance. Overfitting is reduced by setting a max_depth of 3 and a subsample rate of 0.8, which limits tree complexity and introduces randomization into data sampling. Using a min_sample_split value of 2 and a min_sample_leaf value of 1, the model performs well in generalization while capturing detailed patterns. By reducing variance, the usage of max features set to'sqrt' ensures that a subset of features is considered for each split, increasing model robustness. In order to achieve the best possible balance between generalization and accuracy, these hyperparameters were selected using domain expertise and the model's performance on the validation set.

Table 2

GradientBoostingRegressor parameters used for prediction

Parameters	Values
Learning_Rate	0.1
N_Estimators	200
Max_Depth	3
Min_Samples_Split	2
Min_Samples_Leaf	1
Subsamples	0.8
Max_Features	sqrt

The scatter plot in Fig. 3 clearly shows that there is a good correlation between actual and predicted Yield ,with the majority of the data points closely following a predicted straight line .even though there is a wide range of data points between 500 and 3,000 kg/ha, where the models showed higher reliability, for yield exceeds 3,000 kg/ha, there is a significant amount of variability, which indicates some prediction errors. These errors are likely due to a lack of sufficient training data outside the 500 to 3,000 kg/ha range.

Additionally Mean Absolute percentage error (MAPE) is evaluated for evaluating the accuracy of the models and 11% MAPE value is a reasonably good level of success in predicting agricultural yield; although there is still a room for improvement, particularly in reducing prediction errors for higher yield values.

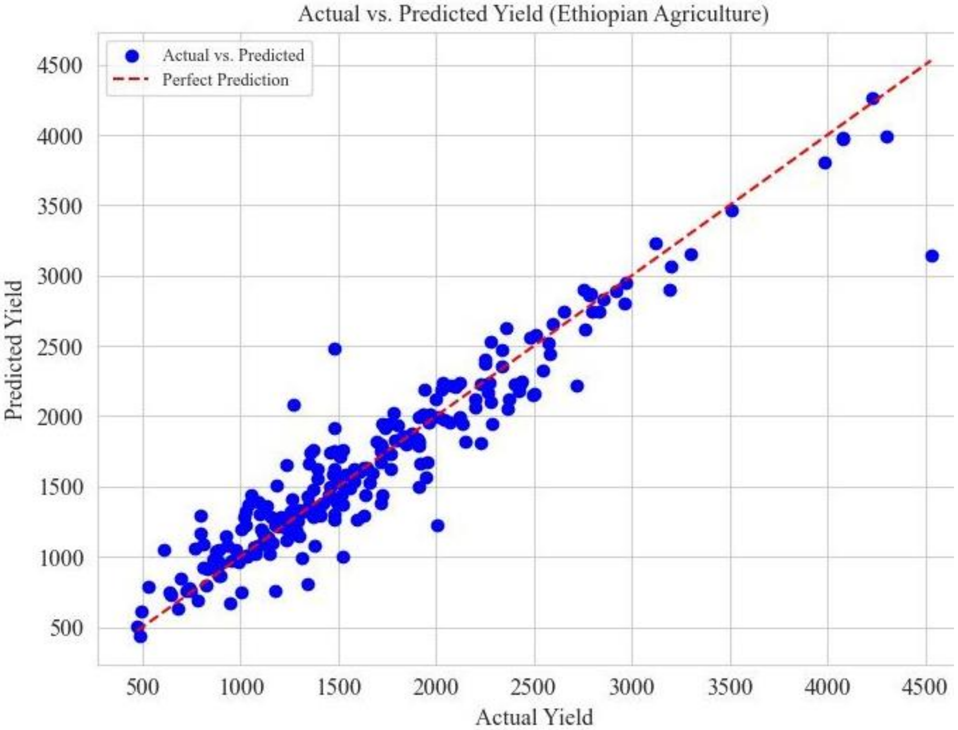


Fig. 3. Actual vs Predicted Yield

Conclusion

There is a lot of promise for increasing agricultural production and food security in Ethiopia through the use of machine learning techniques for crop yield prediction.

The study proposed a GradientBoostingRegressor as the best model for predicting crop yields in Ethiopia based on historical weather, Agricultural and pesticide data. This study achieved RMSE of 0.04, MAE of 0.03, and R2 value of 0.90. DecisionTree regressor also showed good performance characteristics and could also be expanded on further to obtain better results Future studies should evaluate more machine learning models and algorithms on the large and accurate dataset which include other factors like soil quality and irrigation pattern to improve the performance of the models. Additionally, remote sensing data can be merged with statistical data to improve the model's performance. Farmers and policy makers can make informed decisions to optimize crop management practices, reduce risks, and improve overall farming outcomes using historical data and advanced algorithms. Further research and funding in the field of machine language-based crop forecasting is needed to address the various obstacles and prospects that exist in Ethiopia's diverse agricultural sector.

References

1. Agriculture and Food Security. Available at: https://www.usaid.gov/ethiopia/agriculture-and-food-secrity, free (Accessed: July 24, 2024).

2. Taffesse A.S., Dorosh P.A., Asrat S. (2014). Crop production in Ethiopia: Regional Patterns and Trends. In University of Pennsylvania Press eBooks. https://doi.org/10.9783/9780812208610.53.

3. Mekecha B.B. Analyzing climate and agricultural factors for yield prediction of key cereal crops in Ethiopia: A visual analysis (1995–2021). *Modelling and Data Analysis*, 2024, vol. 14 (1), pp. 196–208. DOI: 10.17759/mda.2024140112.

4. Guo Z., Chamberlin J., You L. Smallholder Maize yield estimation using satellite data and machine learning in Ethiopia. *Crop and Environment*, 2023, vol. 2(4), pp. 165–174. DOI:10.1016/j.crope.2023.07.002.

5. Shiferaw H., Getachew T., Sewnet H., Tamene L. Crop Yield Estimation of Teff (Eragrostis tef Zuccagni) Using Geospatial Technology and Machine Learning Algorithm in the Central Highlands of Ethiopia. *Sustainable Agriculture Research*, 2022, vol. 11 (1), p. 34. https://doi.org/10.5539/sar.v11n1p34

6. Araújo S.O. et al. Machine learning applications in agriculture: Current trends, challenges, and future perspectives. *Agronomy*, 2023, vol. 13 (12), p. 2976. DOI: 10.3390/agronomy13122976.

7. Dawn Nabarun, Ghosh Tania, Ghosh Souptik, Saha Aloke, Mukherjee,Pronoy, Sarkar Subhajit, Guha Sagnik, Sanyal Tanmay. Implementation of Artificial Intelligence, Machine Learning, and Internet of Things (IoT) in revolutionizing Agriculture: A review on recent trends and challenges. *International Journal of Experimental Research and Review*, 2023, no. 30, pp. 190–218.

8. Liben F., Abera W., Chernet M.T., Ebrahim M., Tilaye A., Erkossa T., Degefie T.D., Mponela P., Kihara J., Tamene L. Site-specific fertilizer recommendation using data driven machine learning enhanced wheat productivity and resource use efficiency. *Field Crops Research*, 2024, 313 p. 109413.

9. Ayalew A.T., Lohani T.K. Prediction of Crop Yield by Support Vector Machine Coupled with Deep Learning Algorithm Procedures in Lower Kulfo Watershed of Ethiopia. *Journal of Engineering*, 2023, pp. 1–8. https://doi.org/10.1155/2023/6675523.

10. Tefera H.A., Dong-Jun H., Njagi K. Implementation of IoT and machine learning for smart farming monitoring system. *International Journal of Sciences: Basic and Applied Research*, 2020, vol. 52 (1), pp. 6–77.

11. Yang Meijian, Wang Guiling, Wu Shu, Block Paul, Lazin Rehenuma, Alexander Sarah, Lala Jonathan, Haider Muhammad Rezaul, Dokou Zoi, Atsbeha Ezana, Koukoula Marika, Shen Xinyi, Peña Malaquias, Nikolopoulos Efthymios, Bagtzoglou Amvrossios, Anagnostou Emmanouil. Seasonal prediction of crop yields in Ethiopia using an analog approach. *Agricultural and Forest Meteorology*, 2023, vol. 331, p. 109347. DOI: 10.1016/j.agrformet.2023.109347.

12. Mann M., Warner J.M., Malik A. Predicting high-magnitude, low-frequency crop losses using machine learning: an application to cereal crops in Ethiopia. *Climatic Change*, 2019, no. 154 (1–2), pp. 211–227. DOI: 10.1007/s10584-019-02432-7.

13. Central Statistical Agency (CSA). 1996–2022. Agricultural Sample Survey. Vol. 1: Report on Area and Production of Major Crops (Private Peasant Holdings, Meher Season), 1994/95 (1987 E.C.). Statistical Bulletin, Addis Ababa. Available at: https://www.statsethiopia.gov.et/our-survey-reports, free/ (Accessed: November 10, 2024).

14. Food and Agriculture Organization of the United Nations,1996–2022. FAOSTAT statistical database. Available at: http://www.fao.org/faostat/en/#data, free (Accessed: November 10, 2024).

15. World Bank. (1996–2022). Climate Change Knowledge Portal. World Bank. Available at: https://climateknowledgeportal.worldbank.org, free (Accessed: November 10, 2024).

Banchigize B. Mekecha
Postgraduate student, Department of Computer-Aided Engineering and Design, Institute of Information Technologies and Computer Sciences,
University of Science and Technology MISIS
4, Leninskiy pr., Moscow, Russia, 119049
ORCID: 0000-0002-4552-6677
Phone: +7-968-018-28-54
Email: banwoman@gmail.com

Alexander V. Gorbatov
Doctor of Engineering, Professor,
Department of Computer-Aided Engineering and Design,
Institute of Information Technologies and Computer Sciences,
University of Science and Technology MISIS
4, Leninskiy pr., Moscow, Russia, 119049
ORCID: 0000-0002-5061-4831
Phone: +7-926-881-19-73
Email: avgorbatov@mail.ru

Мекеча Б.Б., Горбатов А.В.
Прогнозирование урожайности в Эфиопии
с использованием градиентной регрессии

В настоящее время алгоритмы и методы машинного обучения используются во многих областях исследований для достижения практических и продуктивных решений.

Сельское хозяйство является одной из отраслей, где влияние является значительным, особенно в области прогнозирования урожайности и выбора сельскохозяйственных культур, что имеет решающее значение для обеспечения продовольственной безопасности и совершенствования методов ведения сельского хозяйства. В такой стране, как Эфиопия, где экономика в значительной степени зависит от сельского хозяйства и, в частности, от фермерства в целом, использование возможностей искусственного интеллекта и машинного обучения имеет решающее значение. Однако использование этих технологий в сельском хозяйстве Эфиопии остается ограниченным, главным образом, из-за отсутствия хорошо организованных и цифровых наборов данных и технологических достижений.

Целью данного исследования является повышение точности прогнозирования урожайности сельскохозяйственных культур в Эфиопии и предоставление информации, которая может помочь фермерам и политикам повысить урожайность сельскохозяйственных культур. В этом исследовании была разработана модель прогнозирования урожайности сельскохозяйственных культур на основе исторических данных, которая включает такие факторы, как тип культуры, количество осадков, температура, площадь посева, производство и пестициды.

Среди алгоритмов, рассмотренных в этом исследовании, регрессия с ускорением градиента достигла самого высокого значения R-квадрата – 90% по сравнению с другими, что свидетельствует о его наилучшей прогностической способности. Однако в исследовании также признаются контекстуальные преимущества других алгоритмов, подчеркивая важность выбора моделей, подходящих для конкретных наборов данных и целей. Точность и эффективность сельскохозяйственного планирования и распределения ресурсов в Эфиопии могут быть значительно повышены за счет использования методов машинного обучения для прогнозирования производства сельскохозяйственных культур.

DOI: 10.21293/1818-0442-2024-27-3-125-129

Мекеча Банчигизе Базезев
Аспирантка каф. автоматизированного проектирования
и дизайна (АПД) Института информационных
технологий и компьютерных наук (ИТКН)
Университета науки и технологий МИСИС
Ленинский пр-т, 4, г. Москва, Россия, 119049
ORCID: 0000-0002-4552-6677
Тел.: +7-968-018-28-54
Эл. почта: banwoman@gmail.com,

Горбатов Александр Вячеславович
Д-р техн. наук, проф., зав каф. АПД ИТКН
Университета науки и технологий МИСИС
Ленинский пр-т, 4, г. Москва, Россия, 119049
ORCID: 0000-0002-5061-4831
Тел.: +7-926-881-19-73
Эл. почта: avgorbatov@mail.ru