

서리발생 예측 정확도 향상을 위한 방법 연구

김용석, 최원준*, 심교문, 허지나, 강민구, 조세라

국립농업과학원 기후변화평가과

(2021년 11월 1일 접수; 2021년 12월 16일 수정; 2021년 12월 20일 수락)

Study on Improvement of Frost Occurrence Prediction Accuracy

Yongseok Kim, Wonjun Choi*, Kyo-moon Shim, Jina Hur, Mingu Kang, Sera Jo

Climate change Assessment Division, National Institute of Agricultural Sciences, 166,

Nongsaeangmyeong-ro, Iseo-myeon, Wanju-gun, Jeollabuk-do 55365, Republic of Korea

(Received November 1, 2021; Revised December 16, 2021; Accepted December 20, 2021)

ABSTRACT

In this study, we constructed using Random Forest(RF) by selecting the meteorological factors related to the occurrence of frost. As a result, when constructing a classification model for frost occurrence, even if the amount of data set is large, the imbalance in the data set for development of model has been analyzed to have a bad effect on the predictive power of the model. It was found that building a single integrated model by grouping meteorological factors related to frost occurrence by region is more efficient than building each model reflecting high-importance meteorological factors. Based on our results, it is expected that a high-accuracy frost occurrence prediction model will be able to be constructed as further studies meteorological factors for frost prediction.

Key words: Frost, Random forest, Meteorological factors, Model

I. 서 론

서리는 봄철 과수작물의 꽃이 피기 시작할 무렵 발생하여 꽃봉우리에 피해를 주고 결국 결실을 맺지 못하게 하기 때문에 과수농가는 많은 피해를 입고 있으며, 농작물 수확시기인 가을철에는 수확물의 상품성을 크게 떨어뜨려 큰 피해를 입히기도 한다. 이러한 서리는 일반적으로 날씨가 맑으면서 춥고 바람이 잘 불지 않는 날에 많이 발생한다고 알려져 있으며, 이동성 고기압의 영향을 받아 맑고 바람이 없는 날의 야간에 복사 냉각이 진행되면서 접지층의 대기가 안정되고 기온이 0°C 이하로 하강한 경우 수증기가 승화하여 발생한다고 하였다(Kwon *et al.*, 2008). 또한 서리는 복사

냉각에 의해 역전층이 형성되었을 때 지표층이 상층부보다 차가워진 상태에서 수증기가 유입되어 발생하거나 대류현상에 의해 습윤한 기단에 차가운 공기가 유입되면서 발생한다고 하였다(Temeyer *et al.*, 2003). 그와 함께 서리발생에 관한 다양한 연구가 진행되었는데 Masaki(2021)는 일본을 대상으로 일최저기온과 상대습도, 풍속 등을 이용하여 서리발생 확률 분포에 대한 연구를 수행하였으며, Zeinali *et al.*(2019)은 이란에서 95지점의 서리발생 빈도 조사하여 fuzzy c-approach clustering algorithm을 이용하여 군집분석을 통해 공간적 연관성을 분석하였다. 특히, 서리를 예측하기 위한 연구로 Han *et al.*(2009)는 서리발생 전날의 24시 기온, 전운량, 5일간의 강수량 등의 기상



* Corresponding Author : Wonjun Choi
(ui90@korea.kr)

요소와 판별분석(Discriminant Analysis, DA)을 통해 서리발생을 예측한 연구가 수행된 연구를 진행하였고, Kim *et al.*(2017)은 전운량, 24시 기온, 이슬점온도, 21시 풍속, 평균풍속을 이용하여 서리발생예측 연구를 수행하였다. 그리고 Lee *et al.*(2016)은 최저기온, 최저 초상온도, 이슬점온도, 풍속, 상대습도, 전운량 등의 기상요인들과 로지스틱 회귀(Logistic Regression, LR)모형과 의사결정트리(Decision Tree, DT) 모형을 이용하여 서리발생예측 연구를 수행하였으며, Chun *et al.*(2021)도 최저기온, 초상 최저온도, 상대습도, 풍속, 전운량 등의 기상요소들과 로지스틱 회귀(Logistic Regression, LR)모형, 랜덤포레스트(Random Forest, RF), 장단기 메모리(long-short Term Memory, LSTM)를 이용하여 서리발생예측을 하고 초상일과 만상일을 예측하는 연구를 수행하였다. 또한, Chung *et al.*(2004)은 찬공기 집적량을 계산하여 공간규모의 상세화를 통해 최저기온을 추정하고 -3°C 기준으로 서리발생 경보 시스템 구축에 대한 연구를 수행하였다.

기상요인들을 이용하여 서리발생을 예측하기 위한 방법으로는 서리가 발생한 전날의 기상요인들을 분석하여 다음날의 서리발생 유무를 예측하는 방법과 서리가 발생한 날의 기상요인들을 분석하고 서리발생의 예측을 위해 모형에 사용된 기상인자들의 기상예보값을 이용하여 서리발생 유무를 예측하는 방법이 있을 것이다. 본 연구에서는 서리발생 예측 모형을 구축하기에 앞서, 서리발생과 관련된 기상요인을 분석하고 모형 구축을 위한 데이터 세트의 구성을 분석하여 서리발생 예측모형의 정확도를 좀 더 향상시킬 방안을 모색하기 위해 수행하였다.

II. 재료 및 방법

2.1. 서리가 발생일에 대한 기상자료 수집

2014년~2019년 기간의 24지역을 대상으로 서리가 발생한 날에 대한 자료를 기상청 국립기상과학원으로부터 제공받았으며, 서리가 발생한 날을 기준으로 기상자료개방포털(<https://data.kma.go.kr>)에서 해당 지역의 기상관측자료를 수집하였다. 기상인자는 기존 문헌을 통해 서리발생과 관련이 높을 것으로 예상되는 인자를 중심으로 최저기온, 강수량, 평균풍속, 이슬점온도, 상대습도, 평균 전운량, 평균 지면온도, 최저 초상온도를 선정하였다(Han *et al.*, 2009; Mort and Robinson, 1996; Temeyer *et al.*, 2003).

2.2. 기상인자를 이용한 서리발생일 분류모형

서리가 발생한 날과 서리가 발생하지 않은 날을 분류하기 위해 최저기온, 강수량, 풍속, 이슬점온도, 상대습도, 전운량, 지면온도, 최저 초상온도의 관측데이터를 수집하였다. 분류모형 구축을 위한 방법으로 RF를 이용하였으며, 데이터 세트의 구성을 다르게 하여 여러 번 분류모형을 구축하였다. 첫번째는, 서리가 발생할 확률이 거의 없는 6월~8월의 여름철 자료를 제외하고 나머지 기간에 대해서 서리가 발생한 날에 대해서는 4,424개의 데이터세트로, 서리가 발생하지 않은 날에 대해서는 43,673개의 세트로 구성하였다. RF의 경우 일반적으로 분류모형에 많이 사용되고 있으며, 다수의 DT모형의 결합을 통해 최적의 모형을 형성하는 앙상블기법이다. RF의 기초가 되는 DT는 노드를 분할할 때 불순도함수를 이용하여 분류변수와 분류기준값을 선택하며, 이 과정에서 각 노드에 속하는 자료가 분할하며 자라나게 된다. 여기에 RF는 예측변수들을 임의로 추출하고 추출된 변수 내에서 최적의 분할을 수행하는 방법으로 다수의 DT를 생성해 나간다. 분류모형을 구축하기 위해 통계프로그램 R(4.0.4)을 이용하였으며, RF의 요소별 중요도 분석과 서리발생 유무를 분류 위하여 “randomForest” 패키지의 randomForest 함수를 이용하였다. 좀 더 많은 데이터 세트 확보하여 모형을 검증하기 위해 전체자료를 5개의 집합으로 나누어서 4개의 집합을 훈련자료로 사용하고 나머지 1개의 집합을 검증자료로 사용하면서 이러한 과정을 5번 반복하는 5-fold 교차검증(5-Fold Cross-Validation)을 실시하였다. 두번째는, 서리가 발생하지 않은 날의 데이터 세트 규모가 상대적으로 서리가 발생한 날의 데이터 세트보다 훨씬 크기 때문에 모형의 정확도 분석에 왜곡이 생길 수 있다. 그렇기 때문에 서리가 발생한 날과 서리가 발생하지 않은 날의 학습 데이터 비율을 똑같이 설정하기 위해 서리가 발생한지 않은 날의 데이터 세트는 랜덤으로 추출하여 서리가 발생한 날의 데이터 개수와 동일하게 맞추어 각각 4,424세트씩 구성하였고, 첫번째 방법과 동일하게 k-fold 교차검증 방법을 이용하여 5번 반복 검증과 RF를 이용하여 모형을 구축하였다. 그리고 세번째는, 지역별로 기상요인이 서리 발생에 미치는 영향이 다를 것으로 예상되어, 군집분석(k-means)를 이용하여 24지점에 대한 군집분석을 실시하였고, 그룹화 된 지역을 통합하여 서리가 발생한 날과 발생하지 않은 날에 대해 분류모형을 앞의 방법과 동일하게 구축하였다.

Table 1. Confusion matrix

		Actual	
		Positive	Negative
Predict	Positive	True Positive(TP)	False Positive(FP)
	Negative	False Negative(FN)	True Negative(TN)

2.3. 서리발생 분류모형 평가

서리발생 분류모형에 대한 결과를 Table 1과 같이 정리하고 다음 항목들을 이용하여 평가하였다.

정확도(Accuracy)는 전체 예측한 것에 대해 얼마나 정확하게 예측했는지를 판단하는 지표로, 이 지표를 통해 서리가 발생한 날과 서리가 발생하지 않은 날을 정확히 분류하였는지를 알 수 있다(Eq. 1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (Eq. 1)$$

민감도(Sensitivity)는 실제의 Positive를 잘 예측하

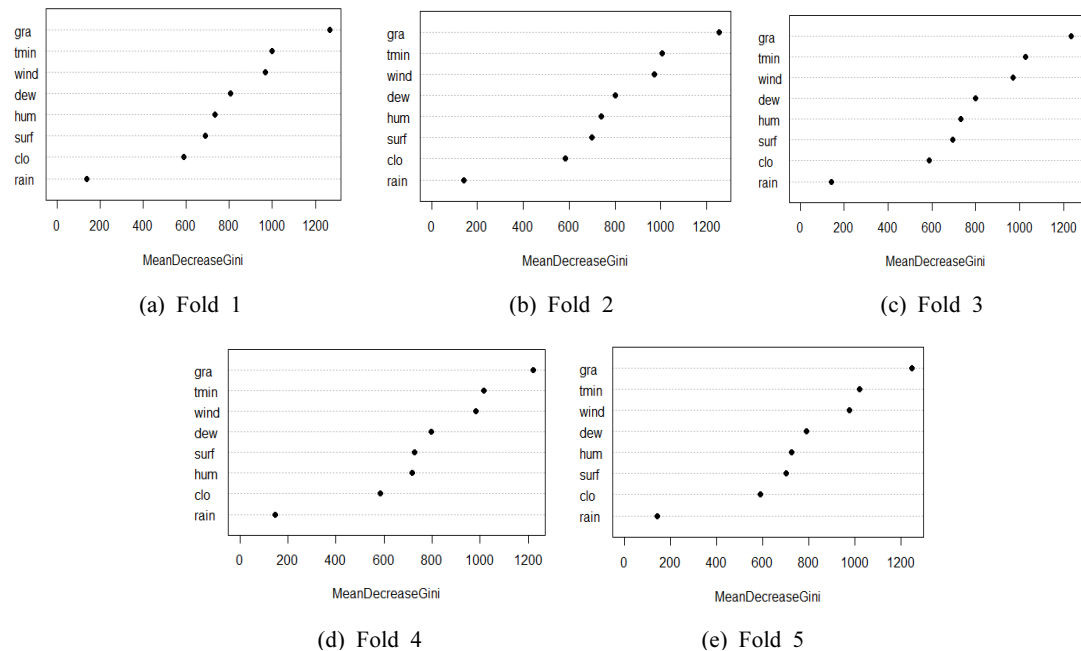
는지를 판단하는 지표로, 이 지표를 통해 서리가 발생한 날에 대해 얼마만큼 서리가 발생한 날로 예측했는지를 알 수 있다(Eq. 2).

$$Sensitivity = \frac{TP}{TP + FN} \quad (Eq. 2)$$

특이도(Specificity)는 실제의 Negative를 잘 예측하는지를 판단하는 지표로, 이 지표를 통해 실제 서리가 발생하지 않은 날에 대해 얼마만큼 서리가 발생하지 않은 날로 예측했는지를 알 수 있다(Eq. 3).

$$Specificity = \frac{TN}{TN + FP} \quad (Eq. 3)$$

균형화 된 정확도(Balanced Accuracy)는 데이터 세트의 불균형 분류문제에서 사용되는 평가지표로, 데이터가 불균형한 상태에서 단순히 정확도(Accuracy)로 평가하기엔 데이터 편향성이 크기 때문에 올바르게 성능을 측정하기 힘든 경우 사용한다. 이는 민감도



Tmin: Minimum temperature, Rain: Precipitation, Wind: Average wind speed, Dew: Average dew point temperature, Clo: Amount of cloud, Hum: Average relative humidity, Surf: Average ground-surface Temperature, Gra: Minimum grass temperature

Fig. 1. Variable importance plot in Random Forest on method 1.

(Sensitivity)와 특이도(Specificity)를 이용하여 계산한다(Eq. 4).

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \quad (Eq.4)$$

III. 결 과

첫번째 방법을 통해서 서리 발생한 날과 서리가 발생하지 않은 날을 분류한 결과 RF의 기상요인에 대한 중요도 분석 결과 초상온도가 가장 중요도가 높게 나타났으며 그 다음으로 최저기온, 풍속 순으로 중요도가 큰 것으로 나타났다(Fig. 1). 이에 반해 전운량과 강수량은 상대적으로 중요도가 작은 것으로 나타났다. 5개의 분류모형의 정확도(Accuracy)는 0.92~0.93으로 평균적으로 0.93의 정확도를 나타냈으며, 균형화

된 정확도(Balanced Accuracy)에서는 0.77~0.80의 정확도를 나타내어 평균적으로 0.79의 정확도를 나타냈다(Table 2).

두번째 방법에서는 학습을 위해 서리가 발생한 날의 데이터 세트와 서리가 발생하지 않은 날의 세트 개수를 동일하게 설정하여 분석을 실시한 결과 서리가 발생한 날의 최저기온 평균은 -3.2°C, 전운량 평균은 3.1할, 지면온도 평균은 2.7°C, 최저 초상온도 평균은 -7.4°C를 나타냈고, 서리가 발생하지 않은 날의 최저기온 평균은 7.5°C, 전운량 평균은 5.0할, 지면온도 평균은 12.9°C, 최저 초상온도 평균은 4.5°C를 나타냈다(Table 3). RF의 기상요인에 대한 중요도 분석(Fig. 2)에서는 최저 초상온도와 최저기온, 지면온도 또는 이슬점온도 순으로 영향력이 큰 것으로 나타나 첫번째 방법의 결과에 비해 풍속의 중요도가 낮은 것을 확인할 수 있었다. 분류모형의 정확도는 0.89~0.91로 나

Table 2. Estimation results of the frost occurrence classification model on method 1

Fold	1		2		3		4		5		Average	
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Actual \ Predict												
Y	532	193	515	222	535	202	561	213	546	192	537.8	204.4
N	352	6116	370	6088	350	6107	324	6097	339	6118	347.0	6105.2
Accuracy	0.92		0.92		0.92		0.93		0.93		0.93	
Sensitivity	0.60		0.58		0.60		0.63		0.62		0.62	
Specificity	0.97		0.96		0.97		0.97		0.97		0.97	
Balanced Accuracy	0.79		0.77		0.79		0.80		0.79		0.79	

Table 3. Characteristic of meteorological factors on frost occurrence day(a) and frost free day(b)

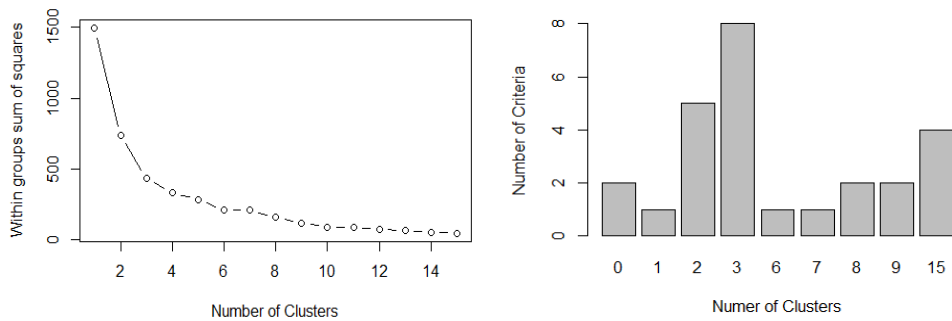
(a)

	Tmin(°C)	Rain(mm)	Wind(m/s)	Dew(°C)	Hum(%)	Clo(0.1%)	Surf(°C)	Gra(°C)
Max	8.4	33.5	13.1	8.2	97.3	10.0	17.0	3.1
Mean	-3.2	0.2	1.5	-5.1	62.3	3.1	2.7	-7.4
Min	-21.7	0.0	0.1	-25.0	18.6	0.0	-8.6	-28.8

(b)

	Tmin(°C)	Rain(mm)	Wind(m/s)	Dew(°C)	Hum(%)	Clo(0.1%)	Surf(°C)	Gra(°C)
Max	25.7	310.0	17.9	25.0	100.0	10.0	36.4	24.4
Mean	7.5	2.7	2.7	4.2	65.0	5.0	12.9	4.5
Min	-21.7	0.0	0.2	-30.0	12.4	0.0	-10.9	-25.4

Tmin: Minimum temperature, Rain: Precipitation, Wind: Average wind speed, Dew: Average dew point temperature, Clo: Amount of cloud, Hum: Average relative humidity, Surf: Average ground-surface Temperature, Gra: Minimum grass temperature



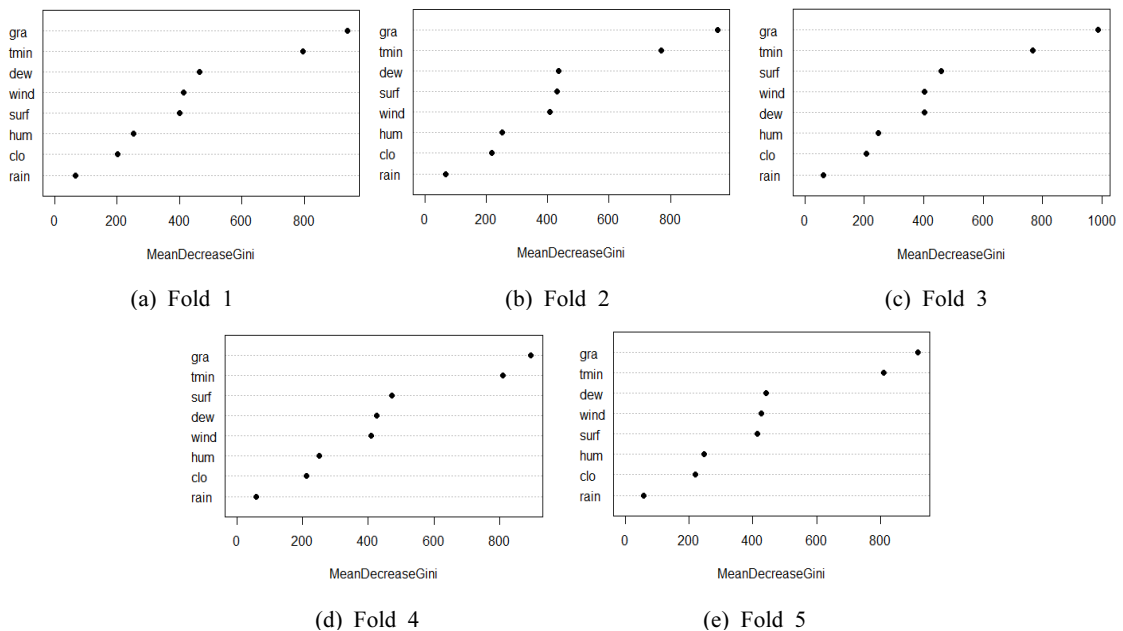
(a) Elbow point

(b) Frequency among 30 indices

Fig. 2. Charts for determining the optimal number of clusters.

Table 4. Estimation results of the frost occurrence classification model on method 2

Fold \ Actual \ Predict	1		2		3		4		5		Average	
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Y	832	134	836	110	825	116	828	130	817	116	827.6	116.0
N	53	750	48	775	60	769	57	755	68	769	68.0	769.0
Accuracy	0.89		0.91		0.90		0.89		0.90		0.90	
Sensitivity	0.94		0.95		0.93		0.94		0.92		0.92	
Specificity	0.85		0.88		0.87		0.85		0.87		0.87	



(a) Fold 1

(b) Fold 2

(c) Fold 3

(d) Fold 4

(e) Fold 5

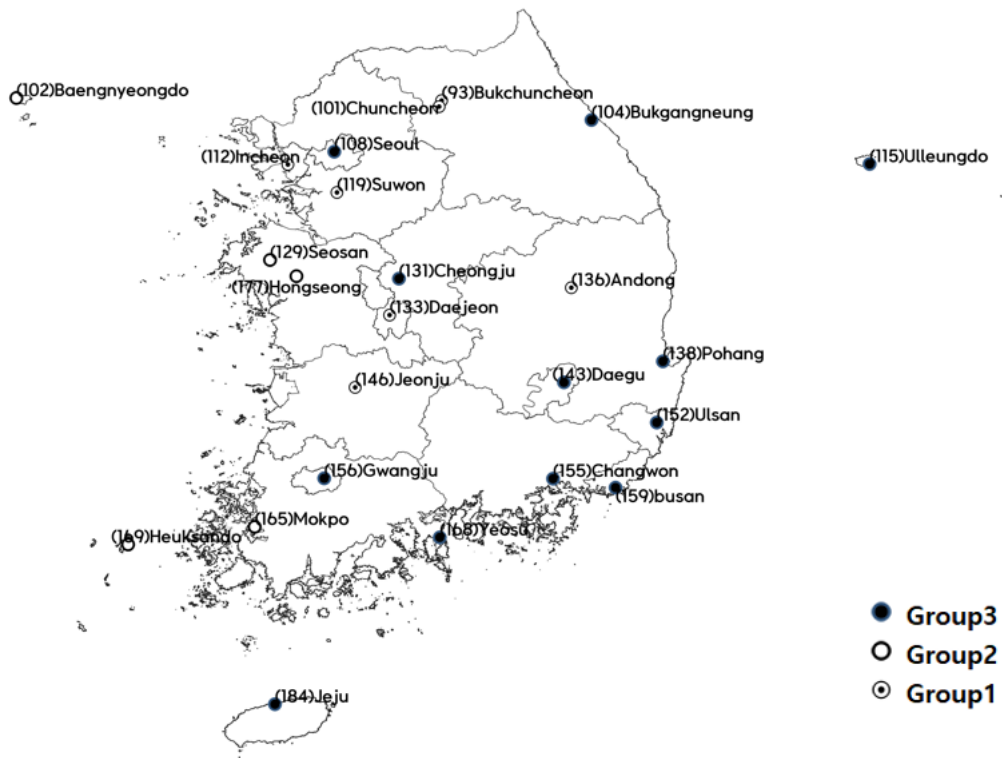
Tmin: Minimum temperature, Rain: Precipitation, Wind: Average wind speed, Dew: Average dew point temperature, Clo: Amount of cloud, Hum: Average relative humidity, Surf: Average ground-surface Temperature, Gra: Minimum grass temperature

Fig. 3. Variable importance plot in Random Forest on method 2.

타났으며, 평균적으로 0.90의 정확도를 나타냈다 (Table 4).

세번째 방법으로 먼저 통계프로그램 R(4.0.4)의 NbClust 함수를 이용하여 25지역을 세리발생과 관련된 기상인자로 적절한 군집의 개수를 분석한 결과 3개로 나누는 것이 가장 알맞을 것으로 나타났다(Fig. 3). 그룹1은 북춘천, 춘천, 인천, 수원, 대전, 안동, 전주가 속하며, 그룹2는 백령도, 서산, 목포, 흑산도, 홍성이 속했다. 그룹3은 북강릉, 서울, 울릉도, 청주, 포항, 대구, 울산, 창원, 광주, 부산, 여수, 제주가 속하였다(Fig. 4). 그룹별 기상특징을 분석한 결과(Table 5), 그룹1은

최저온도, 이슬점온도, 지면온도, 최저 초상온도와 같은 온도관련 요소들이 3개의 그룹 중에서 가장 낮았고, 풍속도 약한 것으로 나타났다. 그룹2는 전반적으로 3개의 그룹 중 풍속이 가장 강했으며, 상대습도도 가장 높았다. 그룹3은 온도관련 요소들이 3개 그룹 중 가장 높았으며, 상대습도는 상대적으로 낮았다. RF의 기상요인에 대한 중요도 분석 결과는 그룹1은 대부분의 세부 모형에서 최저 초상온도, 최저기온, 지면온도 순으로 기상요소의 영향이 큰 것으로 나타났으며(Fig. 5), 그룹2의 경우 대부분의 세부모형에서 최저 초상온도, 최저기온, 이슬점온도의 순으로 영향력이 큰 것



	Group 1	Group 2	Group 3
Station	(93) Bukchuncheon, (101) Chuncheon, (112) Incheon, (119) Suwon, (133) Daejeon, (136) Andong, (146) Jeonju	(102) Baengnyeongdo, (129) Seosan, (165) Mokpo, (169) Heuksando, (177) Hongseong	(104) Bukgangneung, (108) Seoul, (115) Ulleungdo, (131) Cheongju, (138) Pohang, (143) Daegu, (152) Ulsan, (155) Changwon, (156) Gwangju, (159) busan, (168) Yeosu, (184) Jeju

Fig. 4. Result of clustering using k-means.

Table 5. Characteristic of meteorological factors on frost occurrence day by Group

(a) Group 1

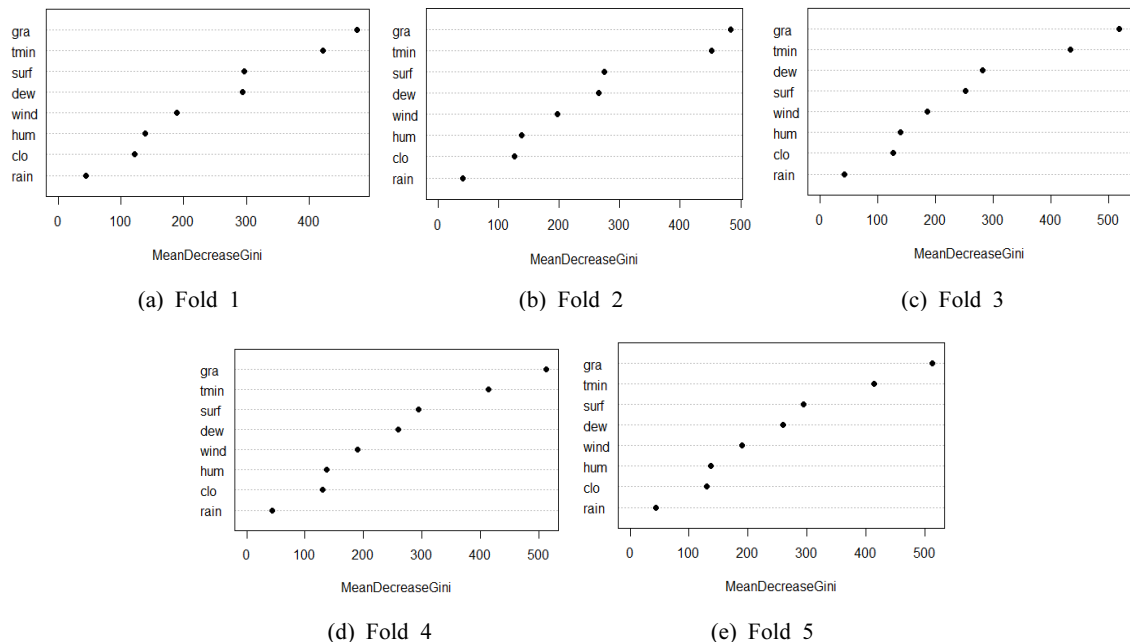
	Tmin(°C)	Rain(mm)	Wind(m/s)	Dew(°C)	Hum(%)	Clo(0.1%)	Surf(°C)	Gra(°C)
Max	-1.7	0.5	2.7	-4.5	66.3	3.6	3.7	-5.7
Mean	-4.2	0.3	1.5	-5.6	63.2	3.2	1.9	-8.5
Min	-6.9	0.1	1.0	-7.8	51.5	2.3	2.2	-11.3

(b) Group 2

	Tmin(°C)	Rain(mm)	Wind(m/s)	Dew(°C)	Hum(%)	Clo(0.1%)	Surf(°C)	Gra(°C)
Max	2.1	0.3	4.0	0.4	74.4	4.1	6.5	-2.1
Mean	-1.5	0.2	2.5	-1.9	72.2	3.3	4.3	-5.2
Min	-4.7	0.1	1.0	-3.6	70.3	2.7	2.8	-8.1

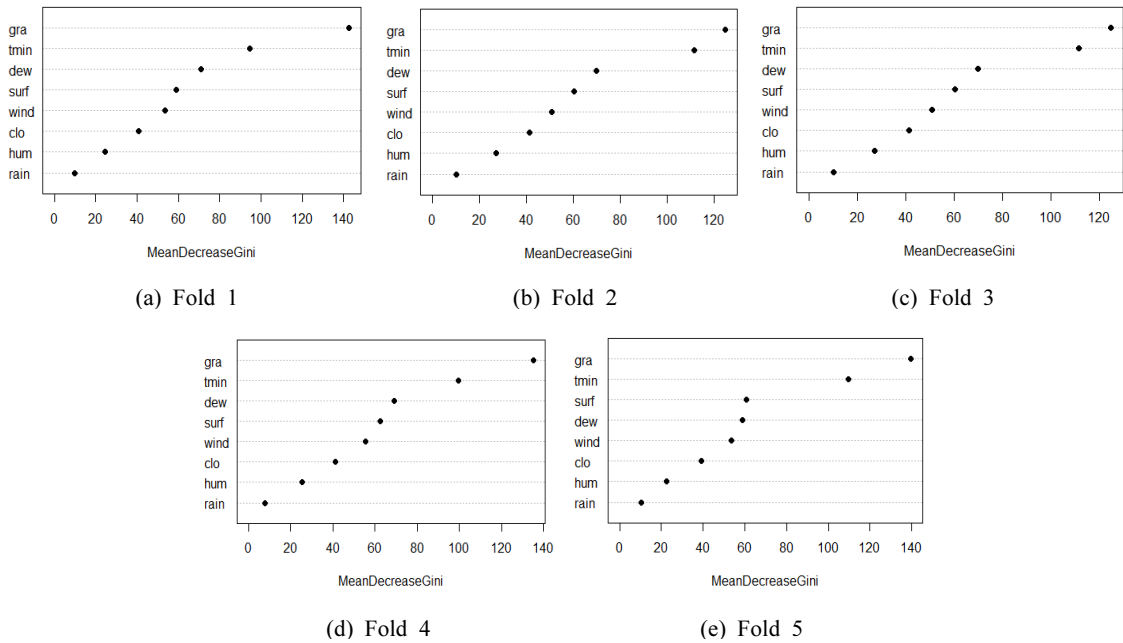
(c) Group 3

	Tmin(°C)	Rain(mm)	Wind(m/s)	Dew(°C)	Hum(%)	Clo(0.1%)	Surf(°C)	Gra(°C)
Max	2.7	0.5	2.9	-0.8	61.9	4.3	6.0	-0.4
Mean	0.0	0.2	1.9	-3.7	57.0	3.0	4.2	-4.7
Min	-2.8	0.0	1.0	-7.8	51.5	2.3	2.2	-7.8



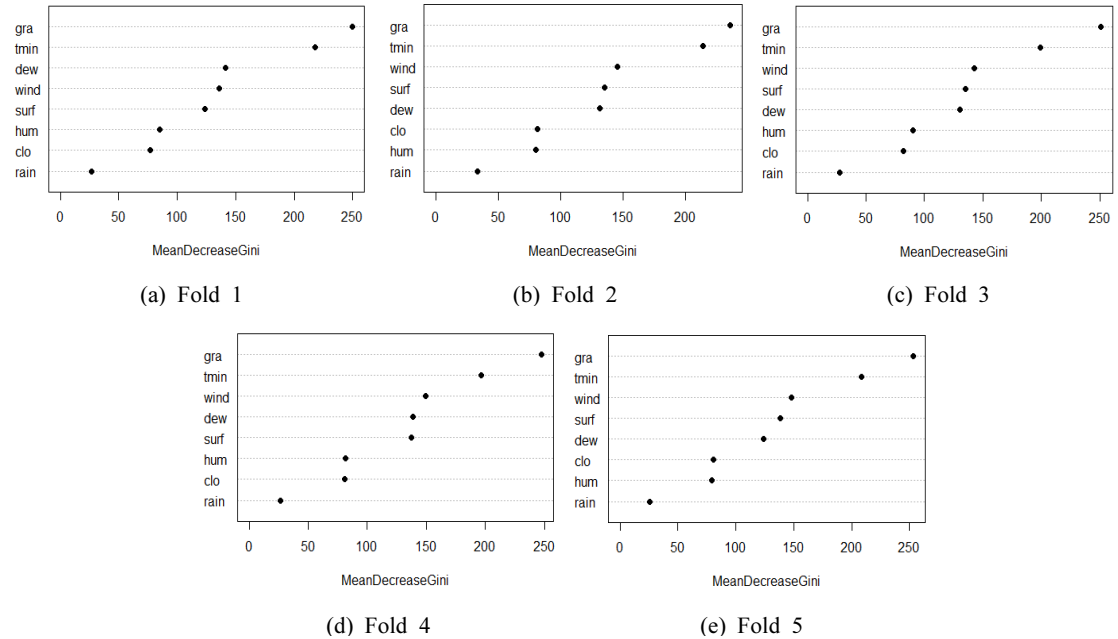
Tmin: Minimum temperature, Rain: Precipitation, Wind: Average wind speed, Dew: Average dew point temperature, Clo: Amount of cloud, Hum: Average relative humidity, Surf: Average ground-surface Temperature, Gra: Minimum grass temperature

Fig. 5. Variable importance plot in Random Forest on method 3(group 1).



Tmin: Minimum temperature, Rain: Precipitation, Wind: Average wind speed, Dew: Average dew point temperature, Clo: Amount of cloud, Hum: Average relative humidity, Surf: Average ground-surface Temperature, Gra: Minimum grass temperature

Fig. 6. Variable importance plot in Random Forest on method 3(group 2).



Tmin: Minimum temperature, Rain: Precipitation, Wind: Average wind speed, Dew: Average dew point temperature, Clo: Amount of cloud, Hum: Average relative humidity, Surf: Average ground-surface Temperature, Gra: Minimum grass temperature

Fig. 7. Variable importance plot in Random Forest on method 3(group 3).

로 나타났다(Fig. 6). 그룹3의 경우 많은 세부모형에서 최저 초상온도, 최저기온, 풍속 순으로 영향력이 큰 것으로 나타났다(Fig. 7). 이에 따라 모든 그룹에서 최저 초상온도와 최저기온이 중요인자로 분석되었지만, 그룹1에서는 지면 온도, 그룹2에서는 이슬점온도, 그룹3에서 풍속이 중요인자로 분석이 되어 지역별 차이가 있는 것으로 분석되었다. 그룹1의 모형의 정확도는 0.86~0.89로 평균적으로 0.88로 나타났다. 그룹2의 경우 0.88~0.93의 정확도를 나타내어 평균적으로 0.90의 정확도를 나타냈다. 그룹3의 모형의 경우 0.87~0.89로 평균적으로 0.88의 정확도를 나타냈다(Table 6).

IV. 결 론

본 연구에서는 기계학습법 중 분류모형에서 많이 사용하고 있는 RF를 이용하여 기상요소에 의한 서리 발생을 정확히 분류할 수 있는지를 분석하였고, 서리 발생과 관련된 기상요인에 대한 중요도분석을 실시하였다.

첫번째 방법에서 학습 데이터 세트의 양이 가장 많아 정확도가 평균적으로 0.93로 가장 높게 나타났지만 모형을 구성할 때 서리가 발생하지 않은 날의 데이터 세트의 비율이 월등히 높았기 때문에 서리가 발생한 날에 비해 발생하지 않은 날의 다양한 기상환경을 더 반영하여 학습될 수 있어 실제로는 서리가 발생하였는데

Table 6. Estimation results for each group of the frost occurrence classification model on method 3

(a) Group 1

Fold \ Actual \ Predict	1		2		3		4		5		Average	
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Y	461	81	465	90	456	84	467	85	456	76	461.0	83.2
N	35	414	40	406	40	412	28	411	40	420	36.6	412.6
Accuracy	0.88		0.87		0.86		0.89		0.88		0.88	
Sensitivity	0.93		0.92		0.92		0.94		0.92		0.93	
Specificity	0.84		0.82		0.83		0.83		0.85		0.83	

(b) Group 2

Fold \ Actual \ Predict	1		2		3		4		5		Average	
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Y	120	14	113	19	116	9	113	17	113	13	115.0	14.4
N	4	110	11	106	8	115	12	107	11	111	9.2	109.8
Accuracy	0.93		0.88		0.93		0.88		0.90		0.90	
Sensitivity	0.97		0.91		0.94		0.90		0.91		0.93	
Specificity	0.89		0.85		0.93		0.86		0.90		0.89	

(c) Group 3

Fold \ Actual \ Predict	1		2		3		4		5		Average	
	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Y	242	47	242	48	247	41	243	35	248	44	244.4	43.0
N	23	217	23	217	18	224	22	230	17	221	20.6	221.8
Accuracy	0.87		0.89		0.89		0.89		0.88		0.88	
Sensitivity	0.91		0.94		0.93		0.92		0.94		0.93	
Specificity	0.82		0.84		0.85		0.87		0.83		0.84	

발생하지 않았다고 예측하는 오류가 높다는 것을 Table 2에서의 민감도가 0.62과 Table 4의 민감도가 0.92인 것을 통해 알 수 있었다. 또한 균형화 된 정확도에서는 0.79로 낮았다. 또한 세번째 방법에서 서리발생과 관련된 기상요소를 이용하여 군집분석을 실시한 후 그룹별로 분석한 결과 최저 초상온도와 최저기온은 모든 그룹에서 중요도가 높게 나타났지만 그룹1에서는 지면 온도, 그룹2에서는 이슬점온도, 그룹3에서 풍속이 상대적으로 중요인자로 분석이 되어 지역적 특징이 나뉘는 것을 볼 수는 있었다. 그렇지만 정확도를 고려한 분류모형을 구축할 경우 그룹별로 영향의 미치는 기상요인의 영향에 비해 모형을 구축하기 위한 전체 데이터 양이 더 중요하게 작용하여 그룹별로 모형을 구축한 모형의 정확도보다 두번째 방법과 같이 지역별 구분 없이 전체의 데이터를 데이터 세트를 구축한 모형이 좀 더 정확도가 높았다. 이것은 학습 데이터 양도 중요하다는 것을 볼 수 있었다. 서리발생과 관련해서 기상요인의 중요도 분석에서는 최저 초상온도와 최저기온, 이슬점온도, 풍속, 지면온도 순으로 서리 발생에 중요한 인자로 분석되었고 강수량이 가장 중요도가 낮은 것으로 나타났다. 이것은 선행 연구에서 서리 발생과 관련인자에 대해 통계적으로 유의성을 분석한 결과에서 최저기온, 풍속, 상대습도, 이슬점온도가 유의했으며, 강수량관련 인자는 유의하지 않았던 것 (Kim et al., 2017)과 어느 정도 일치하는 것을 볼 수 있었다.

이번 연구의 결과에서는 두번째 방법으로 구축된 모형이 가장 효율적인 것으로 나타났다. 그렇지만 본 연구 결과를 바탕으로 서리발생 예측 모형을 개발하기 위해서는 기상예보 자료를 이용하여야 하지만 최저 초상온도나 지면온도와 같이 영향력이 큰 정보는 기상예보로 제공되지 않아 인자로서 활용하기에는 한계가 있을 것이다. 또한 서로 상관성이 높은 온도와 관련인자들이 주로 중요인자로 분석되기 때문에 다중공선성의 문제가 존재할 수 있어 인자 선택에 관한 추가적인 분석이 필요할 것으로 예상된다.

적 요

본 연구에서는 서리발생과 관련된 기상요인을 선정하여 랜덤포레스트(RF)를 이용한 서리발생 유무 분류 모형을 구축하였고, 이와 더불어 기상인자의 중요도와 데이터 세트를 구성하는 방법들을 비교하는 실험을 수

행하였다. 그 결과, 서리발생에 대한 분류 모형을 구축할 경우에 데이터 세트의 양이 많더라도 모형 구축을 위해 학습하기 위한 데이터 세트에서 특정 값이 월등히 많은 불균형은 모형의 예측력에 좋지 못한 영향을 미치는 것으로 분석되었다. 또한, 이번 연구에서 수집된 25지역의 서리발생과 관련된 기상요인에 대해 지역별로 그룹화하여 중요도가 높은 기상요인을 반영한 모형 구축하는 것보다 하나의 통합된 모형을 구축하는 것이 더 효율적인 것으로 나타났다. 이번 연구를 통해 분석된 결과와 서리예측을 위한 기상요인에 대한 추가 분석 연구를 수행한다면 정확도 높은 서리발생 예측 모형을 구축할 수 있을 것이라 예상된다.

감사의 글

이 연구는 농촌진흥청 국립농업과학원 농업과학기술 연구개발사업(과제번호: PJ015151)의 지원으로 수행되었습니다.

REFERENCES

- Chun, J. A., H. J. Lee, S. H. Im, D. Kim, and S. S. Baek, 2021: Comparative assessment of frost event prediction models using logistic regression, random forest, and LSTM networks. *Journal of Korea Water Resources Association* **54**(9), 667-680.
- Chung, U., H. C. Seo, and J. I. Yun, 2004: Site-specific frost warning based on topoclimatic estimation of daily minimum temperature. *Korean Journal of Agricultural and Forest Meteorology* **6**(3), 164-169.
- Han, J. H., J. J. Choi., U. Chung., K. S. Cho, and J. P. Chun, 2009: Frostfall forecasting in the Naju pear production area based on discriminant analysis of climate data. *Korean Journal of Agricultural and Forest Meteorology* **11**(1), 13-18.
- Kim, Y., K. M. Shim, M. P. Jung, and I. T. Choi, 2017: Study on the estimation of frost occurrence classification using machine learning methods. *Korean Journal of Agricultural and Forest Meteorology* **19**(3), 86-92.
- Kwon, Y. A., H. S. Lee., W. T. Kwon, and K. O. Boo, 2008: The weather characteristics of frost occurrence days for protecting crops against frost damage. *Journal of the Korean Geographical Society* **43**(6), 824-842.

- Lee, H., J. A. Chun, H. H. Han, and S. Kim, 2016: Prediction of frost occurrences using statistical modeling approaches. *Advances in Meteorology* **2016**.
- Masaki, Y., 2021: Probability of frost occurrence based on meteorological observation in Japan. *Meteorology and Atmospheric Physics* **133**, 693-705.
- Mort, N., and C. Robinson, 1996: A neural network system for the protection of citrus crops from frost damage. *Computer and Electronics in Agriculture* **16**(3), 177-187.
- Temeyer, B. R., W. A. Gallus Jr, K. A. Jungbluth, D. Burkheimer, and D. McCauley, 2003: Using an artificial neural network to predict parameters for frost deposition on Iowa bridgeways. *Proceedings of the 2003 Mid-Continent Transportation Research Symposium*, Iowa state University, Ames, IA, 2003.
- Zeinali, B., M. Teymouri, S. Asghari, M. Mohammadi, and V. Gupta, 2019: A study of frost occurrence and minimum temperatures in Iran. *Journal of Earth System Science* **128**, 134.